# Appendix A

# The Corpus

## A.1. Format of Article Encoding

```
<!ELEMENT PAPER      (TITLE,REFLABEL,AUTHORS,FILENO,APPEARED,ANNOTATOR?,DATE?,ABSTRACT,
                      BODY,REFERENCES?)>
<!ELEMENT TITLE      (#PCDATA)>
<!ELEMENT AUTHORS    (AUTHOR+)>
<!ELEMENT AUTHOR     (#PCDATA)>
<!ELEMENT FILENO     (#PCDATA)>
<!ELEMENT ANNOTATOR  (#PCDATA)>
<!ELEMENT DATE       (#PCDATA)>
<!ELEMENT YEAR       (#PCDATA)>
<!ELEMENT APPEARED   (#PCDATA)>
<!ELEMENT EQN         EMPTY>
<!ATTLIST EQN
          C           CDATA      'NP'>
<!ELEMENT CREF        EMPTY>
<!ATTLIST CREF
          C           CDATA      'NP'>
<!ELEMENT REFERENCES (P|REFERENCE)*>
<!ELEMENT REFERENCE  (#PCDATA|REFLABEL|W|EQN|NAME|SURNAME|DATE|ETAL|REFAUTHOR|YEAR)*>
<!ELEMENT NAME       (#PCDATA|SURNAME|INVERTED)* >
<!ELEMENT SURNAME    (#PCDATA)>
<!ELEMENT REF        (#PCDATA)*>
<!ATTLIST REF
          SELF       (YES|NO)  "NO"
          C           CDATA      'NNP'>
<!ELEMENT REFAUTHOR  (#PCDATA|SURNAME)*>
<!ATTLIST REFAUTHOR
          C           CDATA     'NNP'>
<!ELEMENT ETAL       (#PCDATA)>
<!ELEMENT BODY       (DIV)+>
<!ELEMENT DIV        (HEADER?, (DIV|P|IMAGE|EXAMPLE)*)>
<!ATTLIST DIV
          DEPTH       CDATA  #REQUIRED >
<!ELEMENT HEADER     (#PCDATA|EQN|REF|REFAUTHOR|CREF|W)*>
<!ATTLIST HEADER      ID  ID  #REQUIRED >
<!ELEMENT P          (S|IMAGE|EXAMPLE)*>
<!ATTLIST P
          TYPE       (ITEM|TXT) "TXT">
<!ELEMENT IMAGE       EMPTY>
<!ATTLIST IMAGE
          ID          ID #REQUIRED
          CATEGORY   (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
```

```
<!ELEMENT S          (#PCDATA|EQN|REF|REFAUTHOR|CREF|FORMULAIC|AGENT|FINITE|W)*>
<!ATTLIST S
          TYPE       (ITEM|TXT) "TXT"
          ID          ID     #REQUIRED
          ABSTRACTC   CDATA  #IMPLIED
          CATEGORY   (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT ABSTRACT   (A-S)*>
<!ELEMENT A-S        (#PCDATA|EQN|REF|REFAUTHOR|CREF|FORMULAIC|AGENT|FINITE|W)*>
<!ATTLIST A-S
          ID          ID         #REQUIRED
          TYPE       (ITEM|TXT)  "TXT"
          DOCUMENTC   CDATA      #IMPLIED
          CATEGORY   (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT EXAMPLE    (EX-S)+>
<!ATTLIST EXAMPLE
          ID          ID #REQUIRED
          CATEGORY   (AIM|CONTRAST|TEXTUAL|OWN|BACKGROUND|BASIS|OTHER)  #IMPLIED>
<!ELEMENT EX-S       (#PCDATA|EQN|W)*>
<!ELEMENT W          (#PCDATA)>
<!ATTLIST W
          C          CDATA  #IMPLIED>
<!ELEMENT FINITE_VERB (#PCDATA)>
<!ATTLIST FINITE_VERB
ACTION
(AFFECT_ACTION|ARGUMENTATION_ACTION|AWARE_ACTION|BETTER_SOLUTION_ACTION|CHANGE_ACTION|
COMPARISON_ACTION|CONTINUE_ACTION|CONTRAST_ACTION|FUTURE_INTEREST_ACTION|INTEREST_ACTION|
NEED_ACTION|PRESENTATION_ACTION|PROBLEM_ACTION|RESEARCH_ACTION|SIMILAR_ACTION|
SOLUTION_ACTION|TEXTSTRUCTURE_ACTION|USE_ACTION|POSSESSION|COPULA|0)
"0">

<!ELEMENT FORMULAIC (#PCDATA|EQN|CREF|REF|REFAUTHOR)*>
<!ATTLIST FORMULAIC TYPE
(US_AGENT|REF_US_AGENT|REF_AGENT|OUR_AIM_AGENT|US_PREVIOUS_AGENT|THEM_PRONOUN_AGENT|THEM_AGENT|
GENERAL_AGENT|PROBLEM_AGENT|SOLUTION_AGENT|THEM_FORMULAIC|US_PREVIOUS_FORMULAIC|
TEXTSTRUCTURE_AGENT|NO_TEXTSTRUCTURE_FORMULAIC|IN_ORDER_TO_FORMULAIC|AIM_FORMULAIC|
TEXTSTRUCTURE_FORMULAIC|METHOD_FORMULAIC|HERE_FORMULAIC|CONTINUE_FORMULAIC|SIMILARITY_FORMULAIC|
COMPARISON_FORMULAIC|CONTRAST_FORMULAIC|GAP_FORMULAIC|FUTURE_FORMULAIC|AFFECT_FORMULAIC|
GOOD_FORMULAIC|BAD_FORMULAIC|0)
"0">

<!ELEMENT AGENT (#PCDATA|EQN|REF|CREF|REFAUTHOR)*>
<!ATTLIST AGENT
  TYPE
  (US_AGENT|THEM_AGENT|THEM_PRONOUN_AGENT|US_PREVIOUS_AGENT|REF_US_AGENT|REF_AGENT|
GENERAL_AGENT|PROBLEM_AGENT|SOLUTION_AGENT|0) "0">
```

## A.2. List of Scientific Articles

| No. | CMP-LG | Conference | Title | Authors | Words | Sent. | Abstr. sent. |
|---|---|---|---|---|---|---|---|
| 0 | 9405001 | ACL94 | Similarity-Based Estimation of Word Cooccurrence Probabilities | I.Dagan, F.Pereira, L.Lee | 4343 | 160 | 7 |
| 1 | 9405002 | ACL94 Student | Temporal Relations: Reference or Discourse Coherence? | A.Kehler | 2320 | 79 | 5 |
| 2 | 9405004 | COLING94 | Syntactic-Head-Driven Generation | E.Koenig | 3438 | 116 | 4 |
| 3 | 9405010 | ACL94 | Common Topics and Coherent Situations: Interpreting Ellipsis in the Context of Discourse Inference | A.Kehler | 5326 | 156 | 5 |
| 4 | 9405013 | COLING94 | Collaboration on Reference to Objects that are not Mutually Known | P.Edmonds | 3994 | 135 | 5 |
| 5 | 9405022 | ACL94 | Grammar Specialization through Entropy Thresholds | C.Samuelsson | 4639 | 170 | 4 |
| 6 | 9405023 | ACL94 Student | An Integrated Heuristic Scheme for Partial Parse Evaluation | A.Lavie | 2454 | 102 | 5 |
| 7 | 9405028 | COLING94 | Semantics of Complex Sentences in Japanese | H.Nakagawa S.Nishizawa | 4700 | 200 | 5 |
| 8 | 9405033 | ACL94 | Relating Complexity to Practical Performance in Parsing with Wide-Coverage Unification Grammars | J.Carroll | 5353 | 121 | 2 |
| 9 | 9405035 | ACL94 Student | Dual-Coding Theory and Connectionist Lexical Selection | Y.Wang | 1889 | 90 | 2 |
| 10 | 9407011 | ACL94 | Discourse Obligations in Dialogue Processing | D.Traum, J.Allen | 6498 | 233 | 2 |
| 11 | 9408003 | COLING94 Reserve | Typed Feature Structures as Descriptions | P.King | 2490 | 167 | 2 |
| 12 | 9408004 | ACL94 Workshop | Parsing with Principles and Probabilities | A.Fordham, M.Crocker | 3645 | 97 | 3 |
| 13 | 9408006 | COLING94 | LHIP: Extended DCGs for Configurable Robust Parsing | A.Ballim, G.Russell | 4468 | 184 | 2 |
| 14 | 9408011 | ACL93 | Distributional Clustering of English Words | F.Pereira, N.Tishby, L.Lee | 4778 | 170 | 4 |
| 15 | 9408014 | ACL94 Workshop | Qualitative and Quantitative Models of Speech Translation | H.Alshawi | 7635 | 296 | 4 |
| 16 | 9409004 | COLING94 | An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus | F.Ribas | 4060 | 179 | 3 |
| 17 | 9410001 | ANLP94 | Improving Language Models by Clustering Training Sentences | D.Carter | 5372 | 150 | 6 |
| 18 | 9410005 | ACL87 | A Centering Approach to Pronouns | S.Brennan, M.Friedman, C.Pollard | 2494 | 98 | 4 |
| 19 | 9410006 | ACL89 | Evaluating Discourse Processing Algorithms | M.Walker | 7281 | 258 | 8 |
| 20 | 9410008 | COLING94 | Recognizing Text Genres with Simple Metrics Using Discriminant Analysis | J.Karlgren, D.Cutting | 1952 | 66 | 3 |
| 21 | 9410009 | COLING94 | Reserve Lexical Functions and Machine Translation | D.Heylen, K.Maxwell, M.Verhagen | 3766 | 135 | 2 |
| 22 | 9410012 | ANLP94 | Does Baum-Welch Re-estimation Help Taggers? | D.Elworthy | 4167 | 1411 | 0 |
| 23 | 9410022 | ACL94 SIG | Automated Tone Transcription | S.Bird | 7139 | 322 | 8 |
| 24 | 9410032 | COLING94 | Planning Argumentative Texts | X.Huang | 3824 | 183 | 4 |
| 25 | 9410033 | COLING94 | Default Handling in Incremental Generation | K.Harbusch, G.Kikui, A.Kilger | 4224 | 176 | 5 |
| 26 | 9411019 | COLING94 | Focus on “only” and “not” | A.Ramsay | 2815 | 99 | 2 |
| 27 | 9411021 | COLING94 | Free-ordered CUG on Chemical Abstract Machine | S.Tojo | 2060 | 86 | 5 |
| 28 | 9411023 | COLING94 | Abstract Generation Based on Rhetorical Structure Extraction | K.Ono, K.Sumita, S.Miike | 2824 | 112 | 4 |

| No. | CMP-LG | Conference | Title | Authors | Words | Sent. | Abstr. sent. |
|---|---|---|---|---|---|---|---|
| 29 | 9412005 | ACL94 SIG | Segmenting Speech without a Lexicon: the Roles of Phonotactics and Speech Source | T.Cartwright, M.Brent | 5481 | 166 | 6 |
| 30 | 9412008 | COLING94 | Analysis of Japanese Compound Nouns using Collocational Information | Y.Kobayasi, T.Tokunaga, H.Tanaka | 3459 | 172 | 4 |
| 31 | 9502004 | COLING94 | Bottom-Up Earley Deduction | G.Erbach | 3591 | 126 | 3 |
| 32 | 9502005 | EACL95 | Off-line Optimization for Earley-style HPSG Processing | G.Minnen, D.Gerdemann, T.Goetz | 4134 | 129 | 3 |
| 33 | 9502006 | EACL95 | Rapid Development of Morphological Descriptions for Full Language Processing Systems | D.Carter | 5292 | 162 | 4 |
| 34 | 9502009 | EACL95 | On Learning More Appropriate Selectional Restrictions | F.Ribas | 3759 | 166 | 4 |
| 35 | 9502014 | EACL95 | Ellipsis and Quantification: A Substitutional Approach | R.Crouch | 5324 | 230 | 2 |
| 36 | 9502015 | EACL95 | The Semantics of Resource Sharing in Lexical-Functional Grammar | A.Kehler, M.Dalrymple, J.Lamping, V.Saraswat | 4259 | 155 | 3 |
| 37 | 9502018 | EACL95 | Algorithms for Analysing the Temporal Structure of Discourse | J.Hitzeman, M.Moens, C.Grover | 3980 | 137 | 4 |
| 38 | 9502021 | EACL95 | A Tractable Extension of Linear Indexed Grammars | B.Keller, D.Weir | 3963 | 140 | 3 |
| 39 | 9502022 | EACL95 | Stochastic HPSG | C.Brew | 3390 | 129 | 3 |
| 40 | 9502023 | EACL95 | Splitting the Reference Time: Temporal Anaphora and Quantification in DRT | R.Nelken, N.Francez | 4283 | 149 | 5 |
| 41 | 9502024 | EACL95 | A Robust Parser Based on Syntactic Information | K.Lee, C.Kweon, J.Seo, G.Kim | 3308 | 159 | 7 |
| 42 | 9502031 | EACL95 Student | Cooperative Error Handling and Shallow Processing | T.Bowden | 2443 | 88 | 6 |
| 43 | 9502033 | EACL95 Student | An Algorithm to Co-Ordinate Anaphora Resolution and PPS Disambiguation Process | S.Azzam | 1301 | 45 | 3 |
| 44 | 9502035 | EACL95 Student | Incorporating “ Unconscious Reanalysis ” into an Incremental, Monotonic Parser | P.Sturt | 4352 | 126 | 4 |
| 45 | 9502037 | EACL95 Student | A State-Transition Grammar for Data-Oriented Parsing | D.Tugwell | 3305 | 116 | 2 |
| 46 | 9502038 | EACL95 Workshop | Implementation and evaluation of a German HMM for POS disambiguation | H.Feldweg | 3625 | 129 | 5 |
| 47 | 9502039 | EACL95 Workshop | Multilingual Sentence Categorization according to Language | E.Giguet | 2142 | 93 | 13 |
| 48 | 9503002 | EACL95 | Computational Dialectology in Irish Gaelic | B.Kessler | 4576 | 165 | 5 |
| 49 | 9503004 | EACL95 Workshop | Creating a Tagset, Lexicon and Guesser for a French tagger | J.Chanod, P.Tapanainen | 4690 | 170 | 3 |
| 50 | 9503005 | EACL95 | A Specification Language for Lexical Functional Grammars | P.Blackburn, C.Gardent | 4968 | 218 | 4 |
| 51 | 9503007 | EACL95 | The Semantics of Motion | P.Sablayrolles | 2361 | 85 | 3 |
| 52 | 9503009 | EACL95 | Distributional Part-of-Speech Tagging | H.Schuetze | 5014 | 184 | 3 |
| 53 | 9503013 | COLING95 | Incremental Interpretation: Applications, Theory, and Relationship to Dynamic Semantics | D.Milward, R.Cooper | 5676 | 186 | 6 |
| 54 | 9503014 | COLING94 | Non-Constituent Coordination: Theory and Practice | D.Milward | 5278 | 192 | 3 |
| 55 | 9503015 | EACL95 | Incremental Interpretation of Categorial Grammar | D.Milward | 4903 | 165 | 4 |

| No. | CMP-LG | Conference | Title | Authors | Words | Sent. | Abstr. sent. |
|---|---|---|---|---|---|---|---|
| 56 | 9503017 | COLING92 | Redundancy in Collaborative Dialogue | M.Walker | 5255 | 212 | 9 |
| 57 | 9503018 | COLING94 | Discourse and Deliberation: Testing a Collaborative Strategy | M.Walker | 5331 | 182 | 4 |
| 58 | 9503023 | EACL95 | A Fast Partial Parse of Natural Language Sentences Using a Connectionist Method | C.Lyon, B.Dickerson | 5027 | 230 | 4 |
| 59 | 9503025 | COLING94 | Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries | Y.Niwa, Y.Nitta | 2749 | 110 | 3 |
| 60 | 9504002 | EACL95 Workshop | Tagset Design and Inflected Languages | D.Elworthy | 3467 | 130 | 3 |
| 61 | 9504006 | ACL88 | Cues and Control in Expert-Client Dialogues | S.Whittaker, P.Stenton | 3925 | 152 | 4 |
| 62 | 9504007 | ACL90 | Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation | M.Walker, S.Whittaker | 5019 | 190 | 9 |
| 63 | 9504017 | ACL95 | A Uniform Treatment of Pragmatic Inferences in Simple and Complex Utterances and Sequences of Utterances | D.Marcu, G.Hirst | 3911 | 132 | 4 |
| 64 | 9504024 | ACL95 | A Morphographemic Model for Error Correction in Nonconcatenative Strings | T.Bowden, G.Kiraz | 3171 | 143 | 4 |
| 65 | 9504026 | ACL95 | The Intersection of Finite State Automata and Definite Clause Grammars | G.vanNoord | 3614 | 151 | 8 |
| 66 | 9504027 | ACL95 | An Efficient Generation Algorithm for Lexicalist MT | V.Poznanski, J.Beaven, P.Whitelock | 4236 | 175 | 3 |
| 67 | 9504030 | ACL95 | Statistical Decision-Tree Models for Parsing | D.Magerman | 4555 | 188 | 8 |
| 68 | 9504033 | ACL95 | Corpus Statistics Meet the Noun Compound: Some Empirical Results | M.Lauer | 4384 | 191 | 4 |
| 79 | 9504034 | ACL95 | Bayesian Grammar Induction for Language Modeling | S.Chen | 4581 | 175 | 5 |
| 70 | 9505001 | ACL95 | Response Generation in Collaborative Negotiation | J.Chu-Carroll, S.Carberry | 5962 | 154 | 5 |
| 71 | 9506004 | ACL95 | Using Higher-Order Logic Programming for Semantic Interpretation of Coordinate Constructs | S.Kulick | 3362 | 130 | 4 |
| 72 | 9511001 | COLING94 | Countability and Number in Japanese-to-English Machine Translation | F.Bond, K.Ogura, S.Ikehara | 3439 | 136 | 2 |
| 73 | 9511006 | ACL95 Workshop | Disambiguating Noun Groupings with Respect to WordNet Senses | P.Resnik | 5970 | 159 | 5 |
| 74 | 9601004 | EACL93 | Similarity between Words Computed by Spreading Activation on an English Dictionary | H.Kozima, T.Furugori | 4384 | 212 | 4 |
| 75 | 9604019 | ACL96 | Magic for Filter Optimization in Dynamic Bottom-up Processing | G.Minnen | 3964 | 157 | 3 |
| 76 | 9604022 | ACL96 | Unsupervised Learning of Word-Category Guessing Rules | A.Mikheev | 6138 | 236 | 4 |
| 77 | 9605013 | COLING96 | Learning Dependencies between Case Frame Slots | H.Li, N.Abe | 4858 | 170 | 8 |
| 78 | 9605014 | COLING96 | Clustering Words with the MDL Principle | H.Li, N.Abe | 4467 | 167 | 5 |
| 79 | 9605016 | ACL96 | Parsing for Semidirectional Lambek Grammar is NP-Complete | J.Doerre | 3060 | 126 | 4 |

# Appendix B

# Example Paper cmp_lg-9408011

## B.1. XML Format


```
<?xml version='1.0'?>
<!DOCTYPE STRUCT-PAPER SYSTEM "/projects/ltg/users/simone/src/dtd/structure.dtd" [
<!ENTITY S "9408011.p">
]>
<STRUCT-PAPER>
<TITLE> Distributional Clustering of English Words </TITLE>
<AUTHORS>
<AUTHOR>Fernando Pereira</AUTHOR>
<AUTHOR>Naftali Tishby</AUTHOR>
<AUTHOR>Lillian Lee</AUTHOR>
</AUTHORS>
<FILENO>9408011</FILENO>
<APPEARED>ACL93</APPEARED>
<ABSTRACT>
<A-S ID='A-0' DOCUMENTC=S-0;S-164> We describe and experimentally evaluate a method for automatically clustering words according
to their distribution in particular syntactic contexts . </A-S>
<A-S ID='A-1'> Deterministic annealing is used to find lowest distortion sets of clusters . </A-S>
<A-S ID='A-2'> As the annealing parameter increases , existing clusters become unstable and subdivide , yielding a hierarchical ``
soft '' clustering of the data . </A-S>
<A-S ID='A-3'> Clusters are used as the basis for class models of word coocurrence , and the models evaluated with respect to
held-out test data . </A-S>
</ABSTRACT>
<BODY>
<DIV DEPTH='1'>
<HEADER ID='H-0'> Introduction </HEADER>
<P>
<S ID='S-0' ABSTRACTC=A-0> Methods for automatically classifying words according to their contexts of use have both scientific and
practical interest . </S>
<S ID='S-1'> The scientific questions arise in connection to distributional views of linguistic ( particularly lexical ) structure
and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives . </S>
<S ID='S-2'> From the practical point of view , word classification addresses questions of data sparseness and generalization in
statistical language models , particularly models for deciding among alternative analyses proposed by a grammar . </S>
</P>
<P>
<S ID='S-3'> It is well known that a simple tabulation of frequencies of certain words participating in certain configurations ,
for example of frequencies of pairs of a transitive main verb and the head noun of its direct object , cannot be reliably used for
comparing the likelihoods of different alternative configurations . </S>
<S ID='S-4'> The problem is that for large enough corpora the number of possible joint events is much larger than the number of
event occurrences in the corpus , so many events are seen rarely or never , making their frequency counts unreliable estimates of
their probabilities . </S>
</P>
<P>
<S ID='S-5'> <REF>Hindle 1990</REF> proposed dealing with the sparseness problem by estimating the likelihood of unseen events
from that of `` similar '' events that have been seen . </S>
<S ID='S-6'> For instance , one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that
direct object for similar verbs . </S>
<S ID='S-7'> This requires a reasonable definition of verb similarity and a similarity estimation method . </S>
<S ID='S-8'> In <REFAUTHOR>Hindle</REFAUTHOR> 's proposal , words are similar if we have strong statistical evidence that they
tend to participate in the same events . </S>
<S ID='S-9'> His notion of similarity seems to agree with our intuitions in many cases , but it is not clear how it can be used
directly to construct word classes and corresponding models of association . </S>
</P>
```

<P>
<S ID='S-10'> Our research addresses some of the same questions and uses similar raw data , but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves .
</S>
<S ID='S-11'> While it may be worthwhile to base such a model on preexisting sense classes <REF>Resnik 1992</REF> , in the work described here we look at how to derive the classes directly from distributional data . </S>
<S ID='S-12'> More specifically , we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w . </S>
<S ID='S-13'> Most other class-based modeling techniques for natural language rely instead on `` hard '' Boolean classes <REF>Brown et al. 1990</REF> . </S>
<S ID='S-14'> Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words , a potentially unreliable source of information as we noted above . </S>
<S ID='S-15'> Our approach avoids both problems . </S>
</P>
<DIV DEPTH='2'>
<HEADER ID='H-1'> Problem Setting </HEADER>
<P>
<S ID='S-16'> In what follows , we will consider two major word classes , <EQN/> and <EQN/> , for the verbs and nouns in our experiments , and a single relation between them , in our experiments relation between a transitive main verb and the head noun of its direct object . </S>
<S ID='S-17'> Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs (v,n) in the required configuration in a training corpus . </S>
<S ID='S-18'> Some form of text analysis is required to collect such a collection of pairs . </S>
<S ID='S-19'> The corpus used in our first experiment was derived from newswire text automatically parsed by <REFAUTHOR>Hindle</REFAUTHOR> 's parser Fidditch <REF>Hindle 1993</REF> . </S>
<S ID='S-20'> More recently , we have constructed similar tables with the help of a statistical part-of-speech tagger <REF>Church 1988</REF> and of tools for regular expression pattern matching on tagged corpora <REF>Yarowsky 1992</REF> . </S>
<S ID='S-21'> We have not yet compared the accuracy and coverage of the two methods , or what systematic biases they might introduce , although we took care to filter out certain systematic errors , for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like `` say '' . </S>
</P>
<P>
<S ID='S-22'> We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs ; the converse problem is formally similar . </S>
<S ID='S-23'> More generally , the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units ( cluster centroids ) and associations between those hidden units . </S>
</P>
<P>
<S ID='S-24'> For the noun classification problem , the empirical distribution of a noun n is then given by the conditional density <EQN/> . </S>
<S ID='S-25'> The problem we study is how to use the <EQN/> to classify the <EQN/> . </S>
<S ID='S-26'> Our classification method will construct a set <EQN/> of clusters and cluster membership probabilities <EQN/> .
</S>
<S ID='S-27'> Each cluster c is associated to a cluster centroid <EQN/> , which is discrete density over <EQN/> obtained by averaging appropriately the <EQN/> . </S>
</P>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-2'> Distributional Similarity </HEADER>
<P>
<S ID='S-28'> To cluster nouns n according to their conditional verb distributions <EQN/> , we need a measure of similarity between distributions . </S>
<S ID='S-29'> We use for this purpose the relative entropy or Kullback-Leibler ( KL ) distance between two distributions . </S>
</P>
<IMAGE ID='I-0'/>
<P>
<S ID='S-30'> This is a natural choice for a variety of reasons , which we will just sketch here . </S>
</P>
<P>
<S ID='S-31'> First of all , <EQN/> is zero just in case p = q , and it increases as the probability decreases that p is the relative frequency distribution of a random sample drawn according to p . </S>
<S ID='S-32'> More formally , the probability mass given by q to the set of all samples of length n with relative frequency distribution p is bounded by <EQN/> <REF>Cover and Thomas 1991</REF> . </S>
<S ID='S-33'> Therefore , if we are trying to distinguish among hypotheses <EQN/> when p is the relative frequency distribution of observations , <EQN/> gives the relative weight of evidence in favor of <EQN/> . </S>
<S ID='S-34'> Furthermore , a similar relation holds between <EQN/> for two empirical distributions p and p ' and the probability that p and p ' are drawn from the same distribution q . </S>
<S ID='S-35'> We can thus use the relative entropy between the context distributions for two words to measure how likely they are to be instances of the same cluster centroid . </S>
</P>
<P>
<S ID='S-36'> From an information theoretic perspective <EQN/> measures how inefficient on average it would be to use a code based on q to encode a variable distributed according to p . </S>
<S ID='S-37'> With respect to our problem , <EQN/> thus gives us the loss of information in using cluster centroid <EQN/> instead of the actual distribution for word <EQN/> when modeling the distributional properties of n . </S>
</P>
<P>
<S ID='S-38'> Finally , relative entropy is a natural measure of similarity between distributions for clustering because its minimization leads to cluster centroids that are a simple weighted average of member distributions . </S>
</P>
<P>
<S ID='S-39'> One technical difficulty is that <EQN/> is not defined when p'(x) = 0 but <EQN/> . </S>
<S ID='S-40'> We could sidestep this problem ( as we did initially ) by smoothing zero frequencies appropriately <REF>Church and Gale 1991</REF> . </S>
<S ID='S-41'> However , this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes . </S>
<S ID='S-42'> It turns out that the problem is avoided by our clustering technique , since it does not need to compute the KL distance between individual word distributions , but only between a word distribution and average distributions , the current cluster centroids , which are guaranteed to be nonzero whenever the word distributions are . </S>

```
<S ID=’S-43’> This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare
individual objects being considered for grouping . </S>
</P>
</DIV>
</DIV>
<DIV DEPTH=’1’>
<HEADER ID=’H-3’> Theoretical Basis </HEADER>
<P>
<S ID=’S-44’> In general , we are interested on how to organize a set of linguistic objects such as words according to the
contexts in which they occur , for instance grammatical constructions or n-grams . </S>
<S ID=’S-45’> We will show elsewhere that the theoretical analysis outlined here applies to that more general problem , but for
now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as
direct objects . </S>
</P>
<P>
<S ID=’S-46’> Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs . </S>
<S ID=’S-47’> The pair coordinates come from two large sets <EQN/> and <EQN/> , with no preexisting topological or metric
structure , and the training data is a sequence S of N independently drawn pairs . </S>
</P>
<IMAGE ID=’I-1’/>
<P>
<S ID=’S-48’> From a learning perspective , this problem falls somewhere in between unsupervised and supervised learning . </S>
<S ID=’S-49’> As in unsupervised learning , the goal is to learn the underlying distribution of the data . </S>
<S ID=’S-50’> But in contrast to most unsupervised learning settings , the objects involved have no internal structure or
attributes allowing them to be compared with each other . </S>
<S ID=’S-51’> Instead , the only information about the objects is the statistics of their joint appearance . </S>
<S ID=’S-52’> These statistics can thus be seem as a weak form of object labelling analogous to supervision . </S>
</P>
<DIV DEPTH=’2’>
<HEADER ID=’H-4’> Distributional Clustering </HEADER>
<P>
<S ID=’S-53’> While clusters based on distributional similarity are interesting on their own , they can also be profitably seen as
a means of summarizing a joint distribution . </S>
<S ID=’S-54’> In particular , we would like to find a set of clusters <EQN/> such that each conditional distribution <EQN/> can
be approximately decomposed as </S>
</P>
<IMAGE ID=’I-2’/>
<P>
<S ID=’S-55’> where <EQN/> is the membership probability of n in c and <EQN/> is v ’s conditional probability given by the
centroid distribution for cluster c . </S>
</P>
<P>
<S ID=’S-56’> The above decomposition can be written in a more symmetric form as </S>
</P>
<IMAGE ID=’I-3’/>
<P>
<S ID=’S-57’> assuming that <EQN/> and <EQN/> coincide . </S>
<S ID=’S-58’> We will take <CREF/> as our basic clustering model . </S>
</P>
<P>
<S ID=’S-59’> To determine this decomposition we need to solve the two connected problems of finding find suitable forms for
the cluster membership and centroid distributions <EQN/> , and of maximizing the goodness of fit between the model distribution
<EQN/> and the observed data . </S>
</P>
<P>
<S ID=’S-60’> Goodness of fit is determined by the model ’s likelihood of the observations . </S>
<S ID=’S-61’> The maximum likelihood ( ML ) estimation principle is thus the natural tool to determine the centroid distributions
<EQN/> . </S>
</P>
<P>
<S ID=’S-62’> As for the membership probabilities , they must be determined solely by the relevant measure of object-to-cluster
similarity , which in the present work is the relative entropy between object and cluster centroid distributions . </S>
<S ID=’S-63’> Since no other information is available , the membership is determined by maximizing the configuration entropy
subject for a fixed average distortion . </S>
<S ID=’S-64’> With the maximum entropy ( ME ) membership distribution , ML estimation is equivalent to the minimization of the
average distortion of the data . </S>
<S ID=’S-65’> The combined entropy maximization entropy and distortion minimization is carried out by a two-stage iterative
process similar to the EM method <REF>Dempster et al. 1977</REF> . </S>
<S ID=’S-66’> The first stage of an iteration is a maximum likelihood , or minimum distortion , estimation of the cluster
centroids given fixed membership probabilities . </S>
<S ID=’S-67’> In the second iteration stage , the entropy of the membership distribution is maximized with a fixed average
distortion . </S>
<S ID=’S-68’> This joint optimization searches for a saddle point in the distortion-entropy parameters , which is equivalent to
minimizing a linear combination of the two known as free energy in statistical mechanics . </S>
<S ID=’S-69’> This analogy with statistical mechanics is not coincidental , and provide us with a better understanding of the
clustering procedure . </S>
</P>
<DIV DEPTH=’3’>
<HEADER ID=’H-5’> Maximum Likelihood Cluster Centroids </HEADER>
<P>
<S ID=’S-70’> For the maximum likelihood argument , we start by estimating the likelihood of the sequence S of N independent
observations of pairs <EQN/> . </S>
<S ID=’S-71’> Using <CREF/> , the sequence ’s model log likelihood is </S>
</P>
<IMAGE ID=’I-4’/>
<P>
<S ID=’S-72’> Fixing the number of clusters ( model size ) <EQN/> , we want to maximize <EQN/> with respect to the distributions
<EQN/> and <EQN/> . </S>
```

```
<S ID=’S-73’> The variation of <EQN/> with respect to these distributions is </S>
</P>
<IMAGE ID=’I-5’/>
<P>
<S ID=’S-74’> with <EQN/> and <EQN/> kept normalized . </S>
<S ID=’S-75’> Using Bayes ’s formula , we have </S>
</P>
<IMAGE ID=’I-6’/>
<P>
<S ID=’S-76’> or </S>
</P>
<IMAGE ID=’I-7’/>
<P>
<S ID=’S-77’> for any c , which we substitute into <CREF/> to obtain </S>
</P>
<IMAGE ID=’I-8’/>
<P>
<S ID=’S-78’> since <EQN/> . </S>
<S ID=’S-79’> This expression is particularly useful when the cluster distributions <EQN/> and <EQN/> are of exponential form ,
precisely what will be provided by the ME step described below . </S>
</P>
<P>
<S ID=’S-80’> At this point we need to specify the clustering model in more detail . </S>
<S ID=’S-81’> In the derivation so far we have treated <EQN/> and <EQN/> symmetrically , corresponding to clusters not of verbs
or nouns but of verb-noun associations . </S>
<S ID=’S-82’> In principle such a symmetric model may be more accurate , but in this paper we will concentrate on asymmetric
models in which cluster memberships are associated to just one of the components of the joint distribution and the cluster centroids
are specified only by the other component . </S>
<S ID=’S-83’> In particular , the model we use in our experiments has noun clusters with cluster memberships determined by <EQN/>
and centroid distributions determined by <EQN/> . </S>
</P>
<P>
<S ID=’S-84’> The asymmetric model simplifies the estimation significantly by dealing with a single component , but it has
the disadvantage that the joint distribution , <EQN/> has two different and not necessarily consistent expressions in terms of
asymmetric models for the two coordinates . </S>
</P>
</DIV>
<DIV DEPTH=’3’>
<HEADER ID=’H-6’> Maximum Entropy Cluster Membership </HEADER>
<P>
<S ID=’S-85’> While variations of <EQN/> and <EQN/> in equation <CREF/> are not independent , we can treat them separately .
</S>
<S ID=’S-86’> First , for fixed average distortion between the cluster centroid distributions <EQN/> and the data <EQN/> ,
we find the cluster membership probabilities , which are the Bayes ’s inverses of the <EQN/> , that maximize the entropy of the
cluster distributions . </S>
<S ID=’S-87’> With the membership distributions thus obtained , we then look for the <EQN/> that maximize the log likelihood l (
S ) . </S>
<S ID=’S-88’> It turns out that this will also be the values of <EQN/> that minimize the average distortion between the
asymmetric cluster model and the data . </S>
</P>
<P>
<S ID=’S-89’> Given any similarity measure <EQN/> between nouns and cluster centroids , the average cluster distortion is </S>
</P>
<IMAGE ID=’I-9’/>
<P>
<S ID=’S-90’> If we maximize the cluster membership entropy </S>
</P>
<IMAGE ID=’I-10’/>
<P>
<S ID=’S-91’> subject to normalization of <EQN/> and fixed <CREF/> , we obtain the following standard exponential forms for the
class and membership distributions </S>
</P>
<IMAGE ID=’I-11’/>
<IMAGE ID=’I-12’/>
<P>
<S ID=’S-92’> where the normalization sums ( partition functions ) are <EQN/> and <EQN/> . </S>
<S ID=’S-93’> Notice that <EQN/> does not need to be symmetric for this derivation , as the two distributions are simply related
by Bayes ’s rule . </S>
</P>
<P>
<S ID=’S-94’> Returning to the log-likelihood variation <CREF/> , we can now use <CREF/> for <EQN/> and the assumption for the
asymmetric model that the cluster membership stays fixed as we adjust the centroids , to obtain </S>
</P>
<IMAGE ID=’I-13’/>
<P>
<S ID=’S-95’> where the variation of [EQn] is now included in the variation of <EQN/> . </S>
</P>
<P>
<S ID=’S-96’> For a large enough sample , we may replace the sum over observations in <CREF/> by the average over <EQN/> . </S>
</P>
<IMAGE ID=’I-14’/>
<P>
<S ID=’S-97’> which , applying Bayes ’s rule , becomes </S>
</P>
<IMAGE ID=’I-15’/>
<P>
<S ID=’S-98’> At the log-likelihood maximum , the variation <CREF/> must vanish . </S>
<S ID=’S-99’> We will see below that the use of relative entropy for similarity measure makes <EQN/> vanish at the maximum as
well , so the log likelihood can be maximized by minimizing the average distortion with respect to the class centroids while class
membership is kept fixed </S>
</P>
```

```
<IMAGE ID='I-16'/>
<P>
<S ID='S-100'> or , sufficiently , if each of the inner sums vanish </S>
</P>
<IMAGE ID='I-17'/>
</DIV>
<DIV DEPTH='3'>
<HEADER ID='H-7'> Minimizing the Average KL Distortion </HEADER>
<P>
<S ID='S-101'> We first show that the minimization of the relative entropy yields the natural expression for cluster centroids
</S>
</P>
<IMAGE ID='I-18'/>
<P>
<S ID='S-102'> To minimize the average distortion <CREF/> , we observe that the variation of the KL distance between noun and
centroid distributions with respect to the centroid distribution <EQN/> , with each centroid distribution normalized by the
Lagrange multiplier <EQN/> , is given by </S>
</P>
<IMAGE ID='I-19'/>
<P>
<S ID='S-103'> Substituting this expression into <CREF/> , we obtain </S>
</P>
<IMAGE ID='I-20'/>
<P>
<S ID='S-104'> Since the <EQN/> are now independent , we obtain immediately the desired centroid expression <CREF/> , which is
the desired weighted average of noun distributions . </S>
</P>
<P>
<S ID='S-105'> We can now see that the variation <EQN/> vanishes for centroid distributions given by <CREF/> , since it follows
from <CREF/> that </S>
</P>
<IMAGE ID='I-21'/>
</DIV>
<DIV DEPTH='3'>
<HEADER ID='H-8'> The Free Energy Function </HEADER>
<P>
<S ID='S-106'> The combined minimum distortion and maximum entropy optimization is equivalent to the minimization of a single
function , the free energy </S>
</P>
<IMAGE ID='I-22'/>
<P>
<S ID='S-107'> where <EQN/> is the average distortion <CREF/> and H is the cluster membership entropy <CREF/> . </S>
</P>
<P>
<S ID='S-108'> The free energy determines both the distortion and the membership entropy through </S>
</P>
<IMAGE ID='I-23'/>
<P>
<S ID='S-109'> with temperature <EQN/> . </S>
</P>
<P>
<S ID='S-110'> The most important property of the free energy is that its minimum determines the balance between the ‘‘
disordering ’’ maximum entropy and ‘‘ ordering ’’ distortion minimization in which the system is most likely to be found . </S>
<S ID='S-111'> In fact the probability to find the system at a given configuration is exponential in F </S>
</P>
<IMAGE ID='I-24'/>
<P>
<S ID='S-112'> so a system is most likely to be found in its minimal free energy configuration . </S>
</P>
</DIV>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-9'> Hierarchical Clustering </HEADER>
<P>
<S ID='S-113'> The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering <REF>Rose et
al. 1990</REF> , in which the number of clusters is determined through a sequence of phase transitions by continuously increasing
the parameter <EQN/> following an annealing schedule . </S>
</P>
<P>
<S ID='S-114'> The higher <EQN/> , the more local is the influence of each noun on the definition of centroids . </S>
<S ID='S-115'> The dissimilarity plays here the role of distortion . </S>
<S ID='S-116'> When the scale parameter <EQN/> is close to zero , the dissimilarities are almost irrelevant , all words
contribute about equally to each centroid , and so the lowest average distortion solution involves just one cluster which is the
average of all word densities . </S>
<S ID='S-117'> As <EQN/> is slowly increased , a point ( phase transition ) is eventually reached which the natural solution
involves two distinct centroids . </S>
<S ID='S-118'> We say then that the original cluster has split into the two new clusters . </S>
</P>
<P>
<S ID='S-119'> In general , if we take any cluster c and a twin c ’ of c such that the centroid <EQN/> is a small random
pertubation of <EQN/> , below the critical <EQN/> at which c splits the membership and centroid reestimation procedure given by
equations <CREF/> and <CREF/> will make <EQN/> and <EQN/> converge , that is , c and c ’ are really the same cluster . </S>
<S ID='S-120'> But with <EQN/> above the critical value for c , the two centroids will diverge , giving rise to two daughters of
c . </S>
</P>
<P>
<S ID='S-121'> Our clustering procedure is thus as follows . </S>
<S ID='S-122'> We start with very low <EQN/> and a single cluster whose centroid is the average of all noun distributions . </S>
<S ID='S-123'> For any given <EQN/> , we have a current set of leaf clusters corresponding to the current free energy ( local )
minimum . </S>
```

```
<S ID='S-124'> To refine such a solution , we search for the lowest <EQN/> which is the critical value for some current leaf
cluster splits . </S>
<S ID='S-125'> Ideally , there is just one split at that critical value , but for practical performance and numerical accuracy
reasons we may have several splits at the new critical point . </S>
<S ID='S-126'> The splitting procedure can then be repeated to achieve the desired number of clusters or model cross-entropy .
</S>
</P>
<IMAGE ID='I-25'/>
</DIV>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-10'> Clustering Examples </HEADER>
<P>
<S ID='S-127'> All our experiments involve the asymmetric model described in the previous section . </S>
<S ID='S-128'> As explained there , our clustering procedure yields for each value of <EQN/> a set <EQN/> of clusters minimizing
the free energy F , and the asymmetric model for <EQN/> estimates the conditional verb distribution for a noun n by </S>
</P>
<IMAGE ID='I-26'/>
<P>
<S ID='S-129'> where <EQN/> also depends on <EQN/> . </S>
</P>
<P>
<S ID='S-130'> As a first experiment , we used our method to classify the 64 nouns appearing most frequently as heads of direct
objects of the verb `` fire '' in one year ( 1988 ) of Associated Press newswire . </S>
<S ID='S-131'> In this corpus , the chosen nouns appear as direct object heads of a total of 2147 distinct verbs , so each noun is
represented by a density over the 2147 verbs . </S>
</P>
<P>
<S ID='S-132'> Figure <CREF/> shows the five words most similar to the each cluster centroid for the four clusters resulting from
the first two cluster splits . </S>
<S ID='S-133'> It can be seen that first split separates the objects corresponding to the weaponry sense of `` fire '' ( cluster 1
) from the ones corresponding to the personnel action ( cluster 2 ) . </S>
<S ID='S-134'> The second split then further refines the weaponry sense into a projectile sense ( cluster 3 ) and a gun sense (
cluster 4 ) . </S>
<S ID='S-135'> That split is somewhat less sharp , possibly because not enough distinguishing contexts occur in the corpus . </S>
</P>
<IMAGE ID='I-27'/>
<P>
<S ID='S-136'> Figure <CREF/> shows the four closest nouns to the centroid of each of a set of hierarchical clusters derived from
verb-object pairs involving the 1000 most frequent nouns in the June 1991 electronic version of Grolier 's Encyclopedia ( 10 million
words ) . </S>
</P>
</DIV>
<DIV DEPTH='1'>
<HEADER ID='H-11'> Model Evaluation </HEADER>
<P>
<S ID='S-137'> The preceding qualitative discussion provides some indication of what aspects of distributional relationships may
be discovered by clustering . </S>
<S ID='S-138'> However , we also need to evaluate clustering more rigorously as a basis for models of distributional relationships
. </S>
<S ID='S-139'> So , far , we have looked at two kinds of measurements of model quality : </S>
<S ID='S-140' TYPE='ITEM'> relative entropy between held-out data and the asymmetric model , and </S>
<S ID='S-141' TYPE='ITEM'> performance on the task of deciding which of two verbs is more likely to take a given noun as direct
object when the data relating one of the verbs to the noun has been witheld from the training data . </S>
</P>
<P>
<S ID='S-142'> The evaluation described below was performed on the largest data set we have worked with so far , extracted from 44
million words of 1988 Associated Press newswire with the pattern matching techniques mentioned earlier . </S>
<S ID='S-143'> This collection process yielded 1112041 verb-object pairs . </S>
<S ID='S-144'> We selected then the subset involving the 1000 most frequent nouns in the corpus for clustering , and randomly
divided it into a training set of 756721 pairs and a test set of 81240 pairs . </S>
</P>
<DIV DEPTH='2'>
<HEADER ID='H-12'> Relative Entropy </HEADER>
<IMAGE ID='I-28'/>
<P>
<S ID='S-145'> Figure <CREF/> plots the average relative entropy of several data sets to asymmetric clustered models of different
sizes , given by </S>
</P>
<IMAGE ID='I-29'/>
<P>
<S ID='S-146'> where <EQN/> is the relative frequency distribution of verbs taking n as direct object in the test set . </S>
<S ID='S-147'> For each critical value of <EQN/> , we show the relative entropy with respect to the asymmetric model based
on <EQN/> of the training set ( set train ) , of randomly selected held-out test set ( set test ) , and of held-out data for a
further 1000 nouns that were not clustered ( set new ) . </S>
<S ID='S-148'> Unsurprisingly , the training set relative entropy decreases monotonically . </S>
<S ID='S-149'> The test set relative entropy decreases to a minimum at 206 clusters , and then starts increasing , suggesting that
larger models are overtrained . </S>
</P>
<P>
<S ID='S-150'> The new noun test set is intended to test whether clusters based on the 1000 most frequent nouns are useful
classifiers for the selectional properties of nouns in general . </S>
<S ID='S-151'> As the figure shows , the cluster model provides over one bit of information about the selectional properties of
the new nouns , but the overtraining effect is even sharper than for the held-out data involving the 1000 clustered nouns . </S>
</P>
</DIV>
<DIV DEPTH='2'>
<HEADER ID='H-13'> Decision Task </HEADER>
<IMAGE ID='I-30'/>
<P>
<S ID='S-152'> We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to
disambiguation in language analysis . </S>
```

```
<S ID=’S-153’> The task consists judging which of two verbs v and v ’ is more likely to take a given noun n as object , when all
occurrences of ( v , n ) in the training set were deliberately deleted . </S>
<S ID=’S-154’> Thus this test evaluates how well the models reconstruct missing data in the verb distribution for n from the
cluster centroids close to n . </S>
</P>
<P>
<S ID=’S-155’> The data for this test was built from the training data for the previous one in the following way , based on a
suggestion by <REF>Dagan et al. 1993</REF> . </S>
<S ID=’S-156’> A small number ( 104 ) of ( v , n ) pairs with a fairly frequent verb ( between 500 and 5000 occurrences ) was
randomly picked , and all occurrences of each pair in the training set were deleted . </S>
<S ID=’S-157’> The resulting training set was used to build a sequence of cluster models as before . </S>
<S ID=’S-158’> Each model was used to decide which of two verbs v and v ’ are more likely to appear with a noun n where the ( v ,
n ) data was deleted from the training set , and the decisions compared with the corresponding ones derived from the original event
frequencies in the initial data set . </S>
<S ID=’S-159’> More specifically , for each deleted pair ( v , n ) and each verb v ’ that occurred with n in the initial data
either at least twice as frequently or at most half as frequently as v , we compared the sign of <EQN/> with that of <EQN/> for
the initial data set . </S>
<S ID=’S-160’> The error rate for each model is simply the proportion of sign disagreements in the selected ( v , n , v ’ )
triples . </S>
<S ID=’S-161’> Figure <CREF/> shows the error rates for each model for all the selected ( v , n , v ’ ) ( all ) and for just
those exceptional triples in which the log frequency ratio of ( n , v ) and ( n , v ’ ) differs from the log marginal frequency
ratio of v and v ’ . </S>
<S ID=’S-162’> In other words , the exceptional cases are those in which predictions based just on the marginal frequencies ,
which the initial one-cluster model represents , would be consistently wrong . </S>
</P>
<P>
<S ID=’S-163’> Here too we see some overtraining for the largest models considered , although not for the exceptional verbs .
</S>
</P>
</DIV>
</DIV>
<DIV DEPTH=’1’>
<HEADER ID=’H-14’> Conclusions </HEADER>
<P>
<S ID=’S-164’ ABSTRACTC=A-0> We have demonstrated that a general divisive clustering procedure for probability distributions can
be used to group words according to their participation in particular grammatical relations with other words . </S>
<S ID=’S-165’> The resulting clusters are intuitively informative , and can be used to construct class-based word coocurrence
models with substantial predictive power . </S>
</P>
<P>
<S ID=’S-166’> While the clusters derived by the proposed method seem in many cases semantically significant , this intuition
needs to be grounded in a more rigorous assessment . </S>
<S ID=’S-167’> In addition to predictive power evaluations of the kind we have already carried out , it might be worth comparing
automatically-derived clusters with human judgements in a suitable experimental setting . </S>
</P>
<P>
<S ID=’S-168’> Moving further in the direction of class-based language models , we plan to consider additional distributional
relations ( for instance , adjective-noun ) and apply the results of clustering to the grouping of lexical associations in
lexicalized grammar frameworks such as stochastic lexicalized tree-adjoining grammars <REF>Schabes 1992</REF> . </S>
</P>
</DIV>
<DIV DEPTH=’1’>
<HEADER ID=’H-15’> Acknowledgments </HEADER>
<P>
<S ID=’S-169’> We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set , the
Fidditch parser and a verb-object structure filter , Mats Rooth for selecting the objects of ‘‘ fire ’’ data set and many
discussions , David Yarowsky for help with his stemming and concordancing tools , and Ido Dagan for suggesting ways of testing
cluster models . </S>
</P>
</DIV>
</BODY>
<REFERENCES>
<REFERENCE>
<REFLABEL>Brown et al 1999</REFLABEL>
Peter F. <SURNAME>Brown</SURNAME>, Vincent J. <SURNAME>Della</SURNAME> <SURNAME>Pietra</SURNAME>, Peter V.
<SURNAME>deSouza</SURNAME>, Jenifer C. <SURNAME>Lai</SURNAME>, and Robert L. <SURNAME>Mercer</SURNAME>. <DATE>1990</DATE>.
Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, pages 283-298, Paris, France, March.
</REFERENCE>
<REFERENCE>
<REFLABEL>Church and Gale 1991</REFLABEL>
Kenneth W. <SURNAME>Church</SURNAME> and William A. <SURNAME>Gale</SURNAME>. <DATE>1991</DATE>. A comparison of the enhanced
Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 5:19-54.
</REFERENCE>
<REFERENCE>
<REFLABEL>Church 1988</REFLABEL>
Kenneth W. <SURNAME>Church</SURNAME>. <DATE>1988</DATE>. A stochastic parts program and noun phrase parser for unrestricted
text. In Proceedings of the Second Conference on Applied Natural Language Processing, pages 136-143, Austin, Texas. Association for
Computational Linguistics, Morristown, New Jersey.
</REFERENCE>
<REFERENCE>
<REFLABEL>Cover 1991</REFLABEL>
Thomas M. <SURNAME>Cover</SURNAME> and Joy A. <SURNAME>Thomas</SURNAME>. <DATE>1991</DATE>. Elements of Information Theory.
Wiley-Interscience, New York, New York.
</REFERENCE>
<REFERENCE>
<REFLABEL>Dagan et al. 1992</REFLABEL>
Ido <SURNAME>Dagan</SURNAME>, Shaul <SURNAME>Markus</SURNAME>, and Shaul <SURNAME>Markovitch</SURNAME>. <DATE>1992</DATE>.
Contextual word similarity and the estimation of sparse lexical relations. Submitted for publication.
</REFERENCE>
```

```
<REFERENCE>
<REFLABEL>Dempster and Rubin 1977</REFLABEL>
A. P. <SURNAME>Dempster</SURNAME>, N. M. <SURNAME>Laird</SURNAME>, and D. B. <SURNAME>Rubin</SURNAME>. <DATE>1977</DATE>.
Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1-38.
</REFERENCE>
<REFERENCE>
<REFLABEL>Duda and Hart 1973</REFLABEL>
Richard O. <SURNAME>Duda</SURNAME> and Peter E. <SURNAME>Hart</SURNAME>. <DATE>1973</DATE>. Pattern Classification and Scene
Analysis. Wiley-Interscience, New York, New York.
</REFERENCE>
<REFERENCE>
<REFLABEL>Hindle 1990</REFLABEL>
Donald <SURNAME>Hindle</SURNAME>. <DATE>1990</DATE>. Noun classification from predicate-argument structures. In 28th Annual
Meeting of the Association for Computational Linguistics, pages 268-275, Pittsburgh, Pennsylvania. Association for Computational
Linguistics, Morristown, New Jersey.
</REFERENCE>
<REFERENCE>
<REFLABEL>Hindle 1993</REFLABEL>
Donald <SURNAME>Hindle</SURNAME>. <DATE>1993</DATE>. A parser for text corpora. In B.T.S. Atkins and A. Zampoli, editors,
Computational Approaches to the Lexicon. Oxford University Press, Oxford, England. <DATE>To appear</DATE>.
</REFERENCE>
<REFERENCE>
<REFLABEL>Resnik 1992</REFLABEL>
Philip <SURNAME>Resnik</SURNAME>. <DATE>1992</DATE>. WordNet and distributional analysis: A class-based approach to lexical
discovery. In AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques, San Jose, California, July.
</REFERENCE>
<REFERENCE>
<REFLABEL>Rose et al. 1990</REFLABEL>
Kenneth <SURNAME>Rose</SURNAME>, Eitan <SURNAME>Gurewitz</SURNAME>, and Geoffrey C. <SURNAME>Fox</SURNAME>.
<DATE>1990</DATE>. Statistical mechanics and phase transitions in clustering. Physical Review Letters, 65(8):945-948.
</REFERENCE>
<REFERENCE>
<REFLABEL>Schabes 1992</REFLABEL>
Yves <SURNAME>Schabes</SURNAME>. <DATE>1992</DATE>. Stochastic lexicalized tree-adjoining grammars. In Proceeedings of the 14th
International Conference on Computational Linguistics, Nantes, France.
</REFERENCE>
<REFERENCE>
<REFLABEL>Yarowsky 1992</REFLABEL>
David <SURNAME>Yarowsky</SURNAME>. <DATE>1992</DATE>. Personal communication.
</REFERENCE>
</REFERENCES>
</STRUCT-PAPER>
```

# B.2. As Published

## DISTRIBUTIONAL CLUSTERING OF ENGLISH WORDS

**Fernando Pereira**
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974
pereira@research.att.com

**Naftali Tishby**
Dept. of Computer Science
Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

**Lillian Lee**
Dept. of Computer Science
Cornell University
Ithaca, NY
llee@cs.cornell.edu



### Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out test data.

### INTRODUCTION

Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example of frequencies of pairs of a transitive main verb and the head noun of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct word classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden *senses classes* and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or *clusters* $c$ with corresponding cluster membership probabilities $p(c|w)$ for each word $w$. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above. Our approach avoids both problems.

#### Problem Setting

In what follows, we will consider two major word classes, $\mathcal{V}$ and $\mathcal{N}$, for the verbs and nouns in our experiments, and a single relation between them, in our experiments relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies $f_{vn}$ of occurrence of particular pairs $(v, n)$ in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, 1992). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any $n$-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster centroids) and associations between those hidden units.

For the noun classification problem, the empirical distribution of a noun $n$ is then given by the conditional density $p_n(v) = f_{vn}/\sum_v f_{vn}$. The problem we study is how to use the $p_n$ to classify the $n \in \mathcal{N}$. Our classification method will construct a set $\mathcal{C}$ of clusters and cluster membership probabilities $p(c|n)$. Each cluster $c$ is associated to a cluster *centroid* $p_c$, which is discrete density over $\mathcal{V}$ obtained by averaging appropriately the $p_n$.

### Distributional Similarity

To cluster nouns $n$ according to their conditional verb distributions $p_n$, we need a measure of similarity between distributions. We use for this purpose the *relative entropy* or *Kullback-Leibler (KL) distance* between two distributions

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad .$$

This is a natural choice for a variety of reasons, which we will just sketch here.[1]

First of all, $D(p \parallel q)$ is zero just in case $p = q$, and it increases as the probability decreases that $p$ is the relative frequency distribution of a random sample drawn according to $p$. More formally, the probability mass given by $q$ to the set of all samples of length $n$ with relative frequency distribution $p$ is bounded by $2^{-nD(p\|q)}$ (Cover and Thomas, 1991). Therefore, if we are trying to distinguish among hypotheses $q_i$ when $p$ is the relative frequency distribution of observations, $D(p \parallel q_i)$ gives the relative weight of evidence in favor of $q_i$. Furthermore, a similar relation holds between $D(p \parallel p')$ for two empirical distributions $p$ and $p'$ and the probability that $p$ and $p'$ are drawn from the same distribution $q$. We can thus use the relative entropy between the context distributions for two words to measure how likely they are to be instances of the same cluster centroid.

From an information theoretic perspective $D(p \parallel q)$ measures how inefficient on average it would be to use a code based on $q$ to encode a variable distributed according to $p$. With respect to our problem, $D(p_n \parallel p_c)$ thus gives us the loss of information in using cluster centroid $p_c$ instead of the actual distribution for word $p_n$ when modeling the distributional properties of $n$.

Finally, relative entropy is a natural measure of similarity between distributions for clustering because its minimization leads to cluster centroids that are a simple weighted average of member distributions.

One technical difficulty is that $D(p \parallel p')$ is not defined when $p'(x) = 0$ but $p(x) > 0$. We could sidestep this problem (as we did initially) by smoothing zero frequencies appropriately (Church and Gale, 1991). However, this is not very satisfactory because one of the goals of our work is precisely to avoid the problems of data sparseness by grouping words into classes. It turns out that the problem is avoided by our clustering technique, since it does not need to compute the KL distance between individual word distributions, but only between a word distribution and average distributions, the current cluster centroids, which are guaranteed to be nonzero whenever the word distributions are. This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.

## THEORETICAL BASIS

In general, we are interested on how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or $n$-grams. We will show elsewhere that the theoretical analysis outlined here applies to that more general problem, but for now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs. The pair coordinates come from two large sets $\mathcal{N}$ and $\mathcal{V}$, with no preexisting topological or metric structure, and the training data is a sequence $S$ of $N$ independently drawn pairs

$$S_i = (n_i, v_i) \qquad 1 \leq i \leq N \ .$$

From a learning perspective, this problem falls somewhere in between unsupervised and supervised learn-

[1] A more formal discussion will appear in our paper *Distributional Clustering*, in preparation.

ing. As in unsupervised learning, the goal is to learn the underlying distribution of the data. But in contrast to most unsupervised learning settings, the objects involved have no internal structure or attributes allowing them to be compared with each other. Instead, the only information about the objects is the statistics of their joint appearance. These statistics can thus be seem as a weak form of object labelling analogous to supervision.

### Distributional Clustering

While clusters based on distributional similarity are interesting on their own, they can also be profitably seen as a means of summarizing a joint distribution. In particular, we would like to find a set of clusters $\mathcal{C}$ such that each conditional distribution $p_n(v)$ can be approximately decomposed as

$$\hat{p}_n(v) = \sum_{c \in \mathcal{C}} p(c|n) p_c(v) \quad ,$$

where $p(c|n)$ is the membership probability of $n$ in $c$ and $p_c(v) = p(v|c)$ is $v$'s conditional probability given by the centroid distribution for cluster $c$.

The above decomposition can be written in a more symmetric form as

$$\begin{aligned} \hat{p}(n, v) &= \sum_{c \in \mathcal{C}} p(c, n) p(v|c) \\ &= \sum_{c \in \mathcal{C}} p(c) p(n|c) p(v|c) \end{aligned} \qquad (1)$$

assuming that $p(n)$ and $\hat{p}(n)$ coincide. We will take (1) as our basic clustering model.

To determine this decomposition we need to solve the two connected problems of finding find suitable forms for the cluster membership and centroid distributions $p(v|c)$, and of maximizing the goodness of fit between the model distribution $\hat{p}(n, v)$ and the observed data

Goodness of fit is determined by the model's likelihood of the observations. The maximum likelihood (ML) estimation principle is thus the natural tool to determine the centroid distributions $p_c(v)$.

As for the membership probabilities, they must be determined solely by the relevant measure of object-to-cluster similarity, which in the present work is the relative entropy between object and cluster centroid distributions. Since no other information is available, the membership is determined by maximizing the configuration entropy subject for a fixed average distortion. With the maximum entropy (ME) membership distribution, ML estimation is equivalent to the minimization of the average distortion of the data. The combined entropy maximization entropy and distortion minimization is carried out by a two-stage iterative process similar to the EM method (Dempster et al., 1977). The first stage of an iteration is a maximum likelihood, or minimum distortion, estimation of the cluster centroids given fixed membership probabilities. In the second iteration stage, the entropy of the membership distribution is maximized with a fixed average distortion. This joint optimization searches for a *saddle point* in the distortion-entropy parameters, which is equivalent to minimizing a linear combination of the two known as *free energy* in statistical mechanics. This analogy with statistical mechanics is not coincidental, and provide us with a better understanding of the clustering procedure.

**Maximum Likelihood Cluster Centroids** For the maximum likelihood argument, we start by estimating the likelihood of the sequence $S$ of $N$ independent observations of pairs $(n_i, v_i)$. Using (1), the sequence's model log likelihood is

$$l(S) = \log \hat{p}(S) = \sum_{i=1}^{N} \log \sum_{c \in \mathcal{C}} p(c) p(n_i|c) p(v_i|c) \quad .$$

Fixing the number of clusters (model size) $|\mathcal{C}|$, we want to maximize $l(S)$ with respect to the distributions $p(n|c)$ and $p(v|c)$. The variation of $l(S)$ with respect to these distributions is

$$\delta l(S) = \sum_{i=1}^{N} \frac{1}{\hat{p}(n_i, v_i)} \sum_{c \in \mathcal{C}} p(c) \begin{pmatrix} p(v_i|c)\delta p(n_i|c) \\ + \\ p(n_i|c)\delta p(v_i|c) \end{pmatrix} \qquad (2)$$

with $p(n|c)$ and $p(v|c)$ kept normalized. Using Bayes's formula, we have [2]

$$p(n_i|c) p(v_i|c) = \frac{p(c|n_i, v_i)}{p(c)} \hat{p}(n_i, v_i) \ ,$$

or

$$\frac{1}{\hat{p}(n_i, v_i)} = \frac{p(c|n_i, v_i)}{p(c) p(n_i|c) p(v_i|c)}$$

for any $c$, which we substitute into (2) to obtain

$$\delta l(S) = \sum_{i=1}^{N} \sum_{c \in \mathcal{C}} p(c|n_i, v_i) \begin{pmatrix} \delta \log p(n_i|c) \\ + \\ \delta \log p(v_i|c) \end{pmatrix} \qquad (3)$$

since $\delta \log p = \delta p / p$. This expression is particularly useful when the cluster distributions $p(n|c)$ and $p(v|c)$

[2] As usual in clustering models (Duda and Hart, 1973), we assume that the model distribution and the empirical distribution are interchangeable at the solution of the parameter estimation equations, since the model is assumed to be able to represent correctly the data at that solution point. In practice, the data may not come exactly from the chosen model class, but the model obtained by solving the estimation equations may still be the closest one to the data.

are of exponential form, precisely what will be provided by the ME step described below.

At this point we need to specify the clustering model in more detail. In the derivation so far we have treated $p(n|c)$ and $p(v|c)$ symmetrically, corresponding to clusters not of verbs or nouns but of verb-noun associations. In principle such a symmetric model may be more accurate, but in this paper we will concentrate on *asymmetric models* in which cluster memberships are associated to just one of the components of the joint distribution and the cluster centroids are specified only by the other component. In particular, the model we use in our experiments has noun clusters with cluster memberships determined by $p(n|c)$ and centroid distributions determined by $p(v|c)$.

The asymmetric model simplifies the estimation significantly by dealing with a single component, but it has the disadvantage that the joint distribution, $p(n, v)$ has two different and not necessarily consistent expressions in terms of asymmetric models for the two coordinates.

**Maximum Entropy Cluster Membership** While variations of $p(n|c)$ and $p(v|c)$ in equation (3 are not independent, we can treat them separately. First, for fixed average distortion between the cluster centroid distributions $p(v|c)$ and the data $p(v|n)$, we find the cluster membership probabilities, which are the Bayes's inverses of the $p(n|c)$, that maximize the entropy of the cluster distributions. With the membership distributions thus obtained, we then look for the $p(v|c)$ that maximize the log likelihood $l(S)$. It turns out that this will also be the values of $p(v|c)$ that minimize the average distortion between the asymmetric cluster model and the data.

Given any similarity measure $d(n, c)$ between nouns and cluster centroids, the average cluster distortion is

$$\langle D \rangle = \sum_{n \in \mathcal{N}} \sum_{c \in \mathcal{C}} p(c|n) d(n, c) \tag{4}$$

If we maximize the cluster membership entropy

$$H = -\sum_{n \in \mathcal{N}} \sum_{c \in \mathcal{C}} p(c|n) \log p(n|c) \tag{5}$$

subject to normalization of $p(n|c)$ and fixed (4), we obtain the following standard exponential forms for the class and membership distributions

$$p(n|c) = \frac{1}{Z_c} \exp -\beta d(n, c) \tag{6}$$

$$p(c|n) = \frac{1}{Z_n} \exp -\beta d(n, c) \tag{7}$$

where the normalization sums (partition functions) are $Z_c = \sum_n \exp -\beta d(n, c)$ and $Z_n = \sum_c \exp -\beta d(n, c)$. Notice that $d(n, c)$ does not need to be symmetric for this derivation, as the two distributions are simply related by Bayes's rule.

Returning to the log-likelihood variation (3), we can now use (6) for $p(n|c)$ and the assumption for the asymmetric model that the cluster membership stays fixed as we adjust the centroids, to obtain

$$\delta l(S) = -\sum_{i=1}^{N} \sum_{c \in \mathcal{C}} p(c|n_i) \delta \beta d(n_i, c) + \delta \log Z_c \tag{8}$$

where the variation of $p(v|c)$ is now included in the variation of $d(n, c)$.

For a large enough sample, we may replace the sum over observations in (8) by the average over $\mathcal{N}$

$$\delta l(S) = -\sum_{n \in N} p(n) \sum_{c \in \mathcal{C}} p(c|n) \delta \beta d(n, c) + \delta \log Z_c$$

which, applying Bayes's rule, becomes

$$\delta l(S) = -\sum_{c \in \mathcal{C}} \frac{1}{p(c)} \sum_{n \in N} p(n|c) \delta \beta d(n, c) + \delta \log Z_c \tag{9}$$

At the log-likelihood maximum, the variation (9) must vanish. We will see below that the use of relative entropy for similarity measure makes $\delta \log Z_c$ vanish at the maximum as well, so the log likelihood can be maximized by minimizing the average distortion with respect to the class centroids while class membership is kept fixed

$$\sum_{c \in \mathcal{C}} \frac{1}{p(c)} \sum_{n \in \mathcal{N}} p(n|c) \delta d(n, c) = 0 \quad ,$$

or, sufficiently, if each of the inner sums vanish

$$\sum_{c \in \mathcal{C}} \sum_{n \in \mathcal{N}} p(n|c) \delta d(n, c) = 0 \tag{10}$$

**Minimizing the Average KL Distortion** We first show that the minimization of the relative entropy yields the natural expression for cluster centroids

$$p(v|c) = \sum_{n \in \mathcal{N}} p(n|c) p(v|n) \tag{11}$$

To minimize the average distortion (10), we observe that the variation of the KL distance between noun and centroid distributions with respect to the centroid distribution $p(v|c)$, with each centroid distribution normalized by the Lagrange multiplier $\lambda_c$, is given by

$$\begin{aligned} \delta d(n, c) &= \delta \begin{pmatrix} -\sum_{v \in \mathcal{V}} p(v|n) \log p(v|c) \\ + \\ \lambda_c (\sum_{v \in \mathcal{V}} p(v|c) - 1) \end{pmatrix} \\ &= \sum_{v \in \mathcal{V}} \left( -\frac{p(v|n)}{p(v|c)} + \lambda_c \right) \delta p(v|c) \quad . \end{aligned}$$

Substituting this expression into (10), we obtain

$$\sum_c \sum_n \sum_v \left( -\frac{p(v|n)p(n|c)}{p(v|c)} + \lambda_c \right) \delta p(v|c) = 0 \quad .$$

Since the $\delta p(v|c)$ are now independent, we obtain immediately the desired centroid expression (11), which is the desired weighted average of noun distributions.

We can now see that the variation $\delta \log Z_c$ vanishes for centroid distributions given by (11), since it follows from (10) that

$$\begin{aligned} \delta \log Z_c &= -\frac{\beta}{Z_c} \sum_n \exp -\beta d(n,c) \delta d(n,c) \\ &= -\beta \sum_n p(n|c) \delta d(x,c) = 0. \end{aligned}$$

**The Free Energy Function** The combined minimum distortion and maximum entropy optimization is equivalent to the minimization of a single function, the *free energy*

$$\begin{aligned} F &= -\frac{1}{\beta} \sum_n \log Z_n \\ &= \langle D \rangle - H/\beta \end{aligned}$$

where $\langle D \rangle$ is the average distortion (4) and $H$ is the cluster membership entropy (5).

The free energy determines both the distortion and the membership entropy through

$$\begin{aligned} \langle D \rangle &= \frac{\partial \beta F}{\partial \beta} \\ H &= -\frac{\partial F}{\partial T} \quad , \end{aligned}$$

with *temperature* $T = \beta^{-1}$.

The most important property of the free energy is that its minimum determines the balance between the "disordering" maximum entropy and "ordering" distortion minimization in which the system is most likely to be found. In fact the probability to find the system at a given configuration is exponential in $F$

$$P \propto \exp -\beta F \quad ,$$

so a system is most likely to be found in its minimal free energy configuration.

## Hierarchical Clustering

The analogy with statistical mechanics suggests a *deterministic annealing* procedure for clustering (Rose et al., 1990), in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter $\beta$ following an *annealing schedule*.

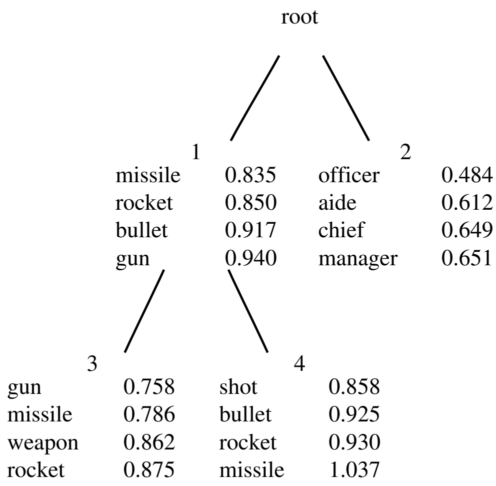



Figure 1: Direct object clusters for *fire*

The higher $\beta$, the more local is the influence of each noun on the definition of centroids. The dissimilarity plays here the role of distortion. When the scale parameter $\beta$ is close to zero, the dissimilarities are almost irrelevant, all words contribute about equally to each centroid, and so the lowest average distortion solution involves just one cluster which is the average of all word densities. As $\beta$ is slowly increased, a point (phase transition) is eventually reached which the natural solution involves two distinct centroids. We say then that the original cluster has *split* into the two new clusters.

In general, if we take any cluster $c$ and a *twin* $c'$ of $c$ such that the centroid $p_{c'}$ is a small random pertubation of $p_c$, below the critical $\beta$ at which $c$ splits the membership and centroid reestimation procedure given by equations (7) and (11) will make $p_c$ and $p_{c'}$ converge, that is, $c$ and $c'$ are really the same cluster. But with $\beta$ above the critical value for $c$, the two centroids will diverge, giving rise to two daughters of $c$.

Our clustering procedure is thus as follows. We start with very low $\beta$ and a single cluster whose centroid is the average of all noun distributions. For any given $\beta$, we have a current set of *leaf* clusters corresponding to the current free energy (local) minimum. To refine such a solution, we search for the lowest $\beta$ which is the critical value for some current leaf cluster splits. Ideally, there is just one split at that critical value, but for practical performance and numerical accuracy reasons we may have several splits at the new critical point. The splitting procedure can then be repeated to achieve the desired number of clusters or model cross-entropy.

## CLUSTERING EXAMPLES

All our experiments involve the asymmetric model described in the previous section. As explained there, our clustering procedure yields for each value of $\beta$ a set $C_\beta$ of clusters minimizing the free energy $F$, and the asymmetric model for $\beta$ estimates the conditional verb distribution for a noun $n$ by

$$\hat{p}_n = \sum_{c \in C_\beta} p(c|n) p_c$$

where $p(c|n)$ also depends on $\beta$.

As a first experiment, we used our method to classify the 64 nouns appearing most frequently as heads of direct objects of the verb "fire" in one year (1988) of Associated Press newswire. In this corpus, the chosen nouns appear as direct object heads of a total of 2147 distinct verbs, so each noun is represented by a density over the 2147 verbs.

Figure 1 shows the five words most similar to the each cluster centroid for the four clusters resulting from the first two cluster splits. It can be seen that first split separates the objects corresponding to the weaponry sense of "fire" (cluster 1) from the ones corresponding to the personnel action (cluster 2). The second split then further refines the weaponry sense into a projectile sense (cluster 3) and a gun sense (cluster 4). That split is somewhat less sharp, possibly because not enough distinguishing contexts occur in the corpus.

Figure 2 shows the four closest nouns to the centroid of each of a set of hierarchical clusters derived from verb-object pairs involving the 1000 most frequent nouns in the June 1991 electronic version of Grolier's Encyclopedia (10 million words).

## MODEL EVALUATION

The preceding qualitative discussion provides some indication of what aspects of distributional relationships may be discovered by clustering. However, we also need to evaluate clustering more rigorously as a basis for models of distributional relationships. So, far, we have looked at two kinds of measurements of model quality: (i) relative entropy between held-out data and the asymmetric model, and (ii) performance on the task of deciding which of two verbs is more likely to take a given noun as direct object when the data relating one of the verbs to the noun has been witheld from the training data.

The evaluation described below was performed on the largest data set we have worked with so far, extracted from 44 million words of 1988 Associated Press newswire with the pattern matching techniques mentioned earlier. This collection process yielded 1112041 verb-object pairs. We selected then the subset involving the 1000 most frequent nouns in the corpus for clustering, and randomly divided it into a training set of 756721 pairs and a test set of 81240 pairs.

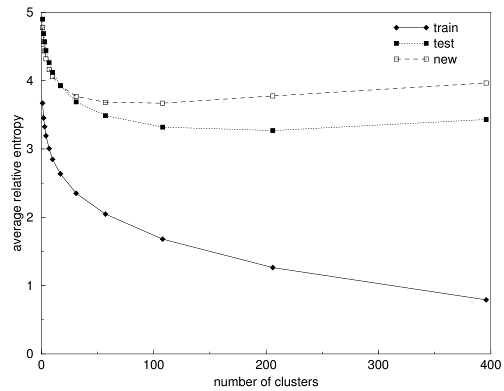



Figure 3: Asymmetric Model Evaluation, AP88 Verb-Direct Object Pairs

### Relative Entropy

Figure 3 plots the average relative entropy of several data sets to asymmetric clustered models of different sizes, given by

$$\sum_n D(t_n || \hat{p}_n)$$

where $t_n$ is the relative frequency distribution of verbs taking $n$ as direct object in the test set. For each critical value of $\beta$, we show the relative entropy with respect to the asymmetric model based on $C_\beta$ of the training set (set *train*), of randomly selected held-out test set (set *test*), and of held-out data for a further 1000 nouns that were not clustered (set *new*). Unsurprisingly, the training set relative entropy decreases monotonically. The test set relative entropy decreases to a minimum at 206 clusters, and then starts increasing, suggesting that larger models are overtrained.

The new noun test set is intended to test whether clusters based on the 1000 most frequent nouns are useful classifiers for the selectional properties of nouns in general. As the figure shows, the cluster model provides over one bit of information about the selectional properties of the new nouns, but the overtraining effect is even sharper than for the held-out data involving the 1000 clustered nouns.

### Decision Task

We also evaluated asymmetric cluster models on a verb decision task closer to possible applications to disambiguation in language analysis. The task consists judging which of two verbs $v$ and $v'$ is more likely to take a

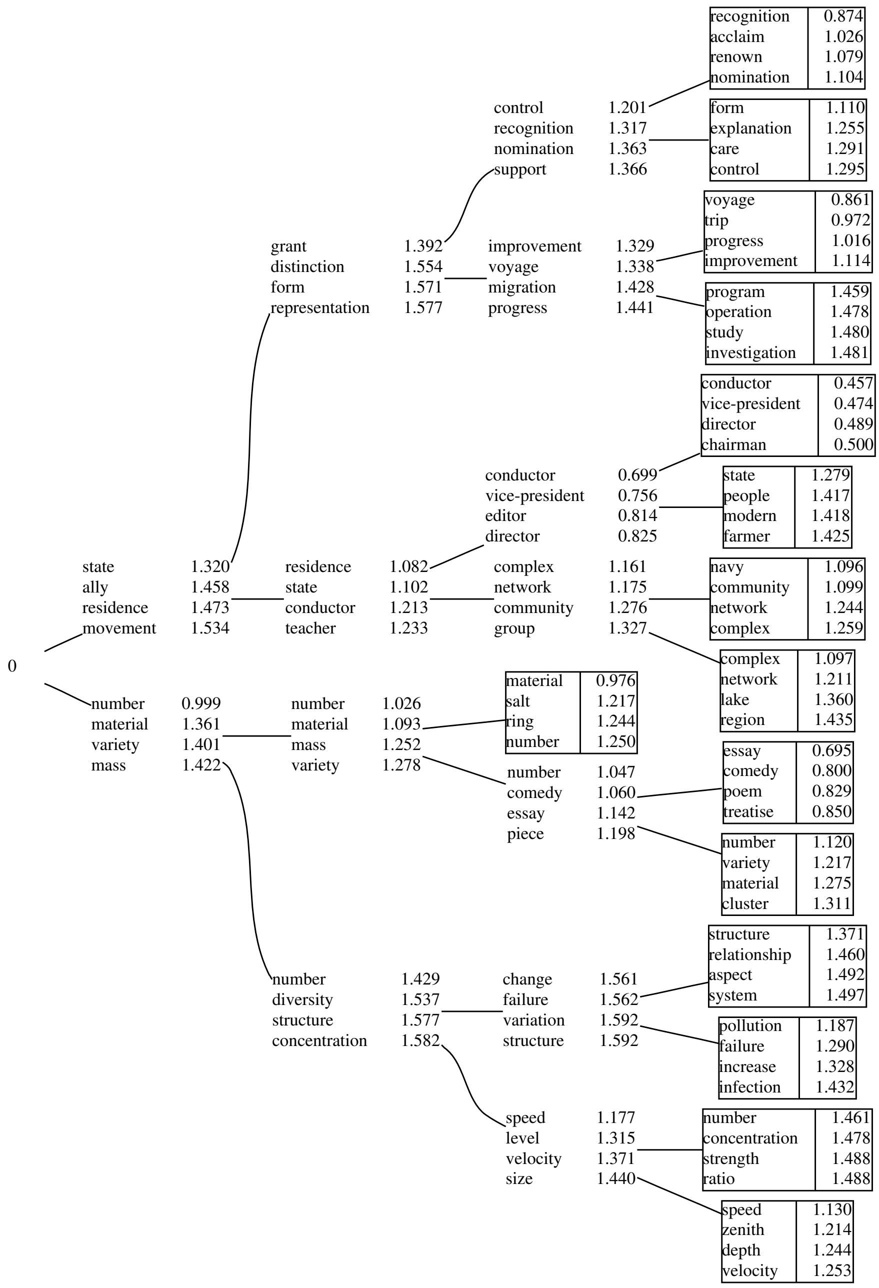



Figure 2: Noun Clusters for Grolier's Encyclopedia

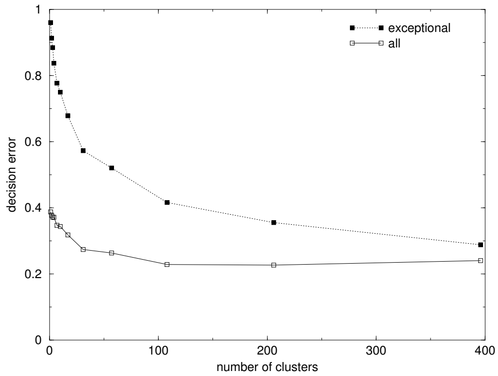



Figure 4: Pairwise Verb Comparisons, AP88 Verb-Direct Object Pairs

given noun $n$ as object, when all occurrences of $(v, n)$ in the training set were deliberately deleted. Thus this test evaluates how well the models reconstruct missing data in the verb distribution for $n$ from the cluster centroids close to $n$.

The data for this test was built from the training data for the previous one in the following way, based on a suggestion by Dagan *et al.* (1992). A small number (104) of $(v, n)$ pairs with a fairly frequent verb (between 500 and 5000 occurrences) was randomly picked, and all occurrences of each pair in the training set were deleted. The resulting training set was used to build a sequence of cluster models as before. Each model was used to decide which of two verbs $v$ and $v'$ are more likely to appear with a noun $n$ where the $(v, n)$ data was deleted from the training set, and the decisions compared with the corresponding ones derived from the original event frequencies in the initial data set. More specifically, for each deleted pair $(v, n)$ and each verb $v'$ that occurred with $n$ in the initial data either at least twice as frequently or at most half as frequently as $v$, we compared the sign of $\log \hat{p}_n(v)/\hat{p}_n(v')$ with that of $\log p_n(v)/p_n(v')$ for the initial data set. The error rate for each model is simply the proportion of sign disagreements in the selected $(v, n, v')$ triples. Figure 4 shows the error rates for each model for all the selected $(v, n, v')$ (*all*) and for just those *exceptional* triples in which the log frequency ratio of $(n, v)$ and $(n, v')$ differs from the log marginal frequency ratio of $v$ and $v'$. In other words, the exceptional cases are those in which predictions based just on the marginal frequencies, which the initial one-cluster model represents, would be consistently wrong.

Here too we see some overtraining for the largest models considered, although not for the exceptional verbs.

## CONCLUSIONS

We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words. The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

While the clusters derived by the proposed method seem in many cases semantically significant, this intuition needs to be grounded in a more rigorous assessment. In addition to predictive power evaluations of the kind we have already carried out, it might be worth comparing automatically-derived clusters with human judgements in a suitable experimental setting.

Moving further in the direction of class-based language models, we plan to consider additional distributional relations (for instance, adjective-noun) and apply the results of clustering to the grouping of lexical associations in lexicalized grammar frameworks such as stochastic lexicalized tree-adjoining grammars (Schabes, 1992).

## ACKNOWLEDGMENTS

We would like to thank Don Hindle for making available the 1988 Associated Press verb-object data set, the Fidditch parser and a verb-object structure filter, Mats Rooth for selecting the objects of "fire" data set and many discussions, David Yarowsky for help with his stemming and concordancing tools, and Ido Dagan for suggesting ways of testing cluster models.

## REFERENCES

[Brown et al.1990] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1990. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, pages 283–298, Paris, France, March.

[Church and Gale1991] Kenneth W. Church and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

[Church1988] Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas. Association for Computational Linguistics, Morristown, New Jersey.

[Cover and Thomas1991] Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory.* Wiley-Interscience, New York, New York.

[Dagan et al.1992] Ido Dagan, Shaul Markus, and Shaul Markovitch. 1992. Contextual word similarity and the estimation of sparse lexical relations. Submitted for publication.

[Dempster et al.1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[Duda and Hart1973] Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, New York.

[Hindle1990] Donald Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania. Association for Computational Linguistics, Morristown, New Jersey.

[Hindle1993] Donald Hindle. 1993. A parser for text corpora. In B.T.S. Atkins and A. Zampoli, editors, *Computational Approaches to the Lexicon.* Oxford University Press, Oxford, England. To appear.

[Resnik1992] Philip Resnik. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-Based Natural-Language-Processing Techniques*, San Jose, California, July.

[Rose et al.1990] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. 1990. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948.

[Schabes1992] Yves Schabes. 1992. Stochastic lexicalized tree-adjoining grammars. In *Proceeedings of the 14th International Conference on Computational Linguistics*, Nantes, France.

[Yarowsky1992] David Yarowsky. 1992. Personal communication.

This paper appeared in the *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190. Ohio State University, Columbus, Ohio, 1993. Reprinted by permission of the Association for Computational Linguistics and the individual authors.

## B.3. RDP

1. SOLUTION IDENTIFIER —

2. SPECIFIC AIM/SCOPE

**164** to group words according to their participation in particular grammatical relations with other words
**10** how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves
**44** how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.
**11** how to derive the classes directly from distributional data
**46** learning a joint distribution of pairs from a large sample of pairs.
**22** we will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs
**45** we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

3. BACKGROUND

| AIM | PROBLEM/PHENOMENON |
|---|---|
| **1** automatically classifying words | **4** The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities. |

4. SOLUTION/INVENTIVE STEP

**164** a general divisive clustering procedure for probability distributions can be used...
**12** we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

5. CLAIM/CONCLUSION

**165** The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

6. RIVAL/CONTRAST

| REFERENCE | SOLUTION ID | TYPE OF CONTRAST |
|---|---|---|
| • **5** [Hindle 1990] | | **9** it is not clear how it can be used directly to construct word classes and corresponding models of association. |
| • **13** [Brown et al. 1992] | **13** other class-based modeling techniques | **13** Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information. |

6. RIVAL/CONTRAST (CT'D)

| | Reference | Solution ID | Type of Contrast |
|---|---|---|---|
| • | **11** [Resnik 1992] | | **11** preexisting sense classes (Resnik) vs. we derive the classes directly from distributional data. |
| • | | **43** agglomerative clustering techniques | **43** need to compare individual objects being considered for grouping. (advantage of our method) |
| • | **40** [Church and Gale 1991] | **40** smoothing zero frequencies appropriately | **41** However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together |

7. BASIS/CONTINUATION

| | Reference | Solution ID | Type of Continuation |
|---|---|---|---|
| • | **113** [Rose et al. 1990] | **113** deterministic annealing | **113** The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990] ... |
| • | **155** [Dagan et al. 1993] | | **155** based on a suggestion by |
| • | | **29** Kullback-Leibler (KL) distance | **29** used |
| • | **19** [Hindle 1993] | | **19** automatically parsed by Hindle's parser |
| • | **20** [Church 1988] | | **20** with the help of a statistical part-of-speech tagger |
| • | **20** [Yarowsky 1992] | | **20** [with the help of] tools for regular expression pattern matching on tagged corpora |

EXTERNAL STRUCTURE

| Headlines | | 8. Textual Structure |
|---|---|---|
| **1.** | Introduction | |
| **1.1** | Problem Setting | |
| **1.2** | Distributional Similarity | |
| **2.** | Theoretical Basis | |
| **2.1** | Distributional Clustering | |
| **2.1.1.** | Maximum Likelihood Cluster Centroids | |
| **2.1.2.** | Maximum Entropy Cluster Membership | |
| **2.1.3.** | Minimizing the Average KL Distortion | |
| **2.1.4.** | The Free Energy Function | |
| **2.2.** | Hierarchical Clustering | |
| **3.** | Clustering Examples | **127** All our experiments involve the asymmetric model described in the previous section. |
| **4.** | Model Evaluation | |
| **4.1.** | Relative Entropy | |
| **4.2.** | Decision Task | |
| **5.** | Conclusions | |

## B.4. RDP Sentence Material

SPECIFIC AIM/SCOPE

**10** Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves.

**11** While it may be worthwhile to base such a model on preexisting sense classes [Resnik 1992], in the work described here we look at how to derive the classes directly from distributional data.

**22** We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar.

**44** In general, we are interested on how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.

**45** We will show elsewhere that the theoretical analysis outlined here applies to that more general problem, but for now we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects.

**46** Our problem can be seen as that of learning a joint distribution of pairs from a large sample of pairs.

**164** We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.

BACKGROUND (AIM)

**1** Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.

BACKGROUND (PROBLEM/PHENOMENON)

**4** The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

SOLUTION/INVENTIVE STEP

**12** More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

**164** We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.

CLAIM/CONCLUSION

**165** The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

RIVAL/CONTRAST

**5** [Hindle 1990] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of “similar” events that have been seen.

**9** His notion of similarity seems to agree with our intuitions in many cases, but is not clear how it can be used directly to construct word classes and corresponding models of association.

**11** While it may be worthwhile to base such a model on preexisting sense classes [Resnik 1992], in the work described here we look at how to derive the classes directly from distributional data.

**13** Most other class-based modeling techniques for natural language rely instead on “hard” Boolean classes [Brown et al. 1990].

**14** Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information as we noted above.

**40** We could sidestep this problem (as we did initially) by smoothing zero frequencies appropriately [Church and Gale 1991].

**41** However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together.

**43** This is a useful advantage of our method compared with agglomerative clustering techniques that need to compare individual objects being considered for grouping.

BASIS/CONTINUATION

**19** The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle’s parser Fidditch [Hindle 1993].

**20** More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger [Church 1988] and of tools for regular expression pattern matching on tagged corpora [Yarowsky 1992].

**29** We use for this purpose the relative entropy or Kullback-Leibler (KL) distance between two distributions.

**113** The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990], in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN> following an annealing schedule.

**155** The data for this test was built from the training data for the previous one in the following way, based on a suggestion by [Dagan et al. 1993].

TEXTUAL STRUCTURE

**127** All our experiments involve the asymmetric model described in the previous section.

## B.5. Human Annotation (Annotator A)

# Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

## Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

## Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

## B.6. Human Annotation (Annotator B)

# Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

### Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

### Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

### Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

## B.7. Agent and Action Recognition

# Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee


### Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.


### Introduction

Methods for automatically classifying words according to
1 their contexts of use have both scientific and practial inte-
2 rest. The scientific questions arise in connection to distri-
butional views of linguistic (particularly lexical) structure
and also in relation to the question of lexical acquisition
both from psychological and computational learning
perspectives. From the practical point of view, word
classification addresses questions of data sparseness and
generalization in statistical language models, particularly
3 models for deciding among alternative analyses proposed
by a grammar.

It is well known that a simple tabulation of frequencies
of certain words participating in certain configurations, for
example the frequencies of pairs of transitive main verb
4 and the head of its direct object, cannot be reliably used
for comparing the likelihoods of different alternative confi-
1 5 gurations. The problem is that in large enough corpora, the
number of possible joint events is much larger than the
number of event occurrences in the corpus, so many events
6 are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

2 7 Hindle (1990) proposed dealing with the sparseness
problem by estimating the likelihood of unseen events from
that of "similar" events that have been seen. For instance,
8 one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for
9 similar verbs. This requires a reasonable definition of verb
similarity and a similarity estimation method. In Hindle's
3 10 proposal, words are similar if we have strong statistical
4 evidence that they tend to participate in the same events.
5 His notion of similarity seems to agree with our intuitions
11 in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and 6 12
uses similar raw data, but we investigate how to factor word 7 13
association tendencies into associations of words to certain
hidden senses classes and associations between the classes
themselves. While it may be worthwhile to base such a model
on preexisting sense classes (Resnik, 1992), in the work descri- 8 14
bed here we look at how to derive the classes directly from 9 15
distributional data. More specifically, we model senses as 10 16
probabilistic concepts or clusters c with corresponding cluster
membership probabilities <EQN/> for each word w. Most
other class-based modeling techniques for natural language
rely instead on "hard" Boolean classes (Brown et al., 1990).
Class construction is then combinatorically very demanding 17
and depends on frequency counts for joint events involving
particular words, a potentially unreliable source of inform-
ation, as we noted above. Our approach avoids both problems. 11 18 12 19

### Problem Setting

In what follows, we will consider two major word classes, 13 20
<EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb
and the head noun of its direct object. Our raw knowledge
about the relation consists of the frequencies <EQN/> of
occurrence of particular pairs <EQN/> in the required con-
figuration in a training corpus. Some form of text analysis
is required to collect such a collection of pairs. The corpus 21
used in our first experiment was derived from newswire text 22 23
automatically parsed by Hindle's parser Fidditch (Hindle,
1993). More recently, we have constructed similar tables 14 24
with the help of a statistical part-of-speech tagger (Church,
1988) and of tools for regular expression pattern matching
on tagged corpora (Yarowsky, p.c.). We have not yet 15
compared the accuracy and coverage of the two methods,
or what systematic biases they might introduce, although 16 25
we took care to filter out certain systematic errors, for in- 17
stance the misparsing of the subject of a complement clause
as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classi- 18 26
fying nouns according to their distribution as direct objects
of verbs; the converse problem is formally similar. More
generally, the theoretical basis for our method supports the 19
use of clustering to build models for any n-ary relation in
terms of associations between elements in each coordinate
and appropriate hidden units (cluster controids) and associations between these hidden units.

| Actions (blue) | |
|---|---|
| 1 | POSSESSION_ACTION |
| 2 | PROBLEM_ACTION |
| 3 | SOLUTION_ACTION (POS-error) |
| 4 | negated USE_ACTION (passive) |
| 5 | COPULA |
| 6 | RESEARCH_ACTION (POS-error) |
| 7 | PRESENTATION_ACTION |
| 8 | RESEARCH_ACTION |
| 9 | NEED_ACTION |
| 10 | POSSESSION_ACTION |
| 11 | USE_ACTION (passive) |
| 12 | INTEREST_ACTION |
| 13 | RESEARCH_ACTION |
| 14 | PRESENTATION_ACTION (POS-error) |
| 15 | INTEREST_ACTION |
| 16 | SOLUTION_ACTION |
| 17 | COPULA |
| 18 | PRESENTATION_ACTION |
| 19 | SOLUTION_ACTION |
| 20 | INTEREST_ACTION |
| 21 | NEED_ACTION |
| 22 | USE_ACTION (POS-error) |
| 23 | CONTINUE_ACTION |
| 24 | RESEARCH_ACTION |
| 25 | PRESENTATION_ACTION |
| 26 | INTEREST_ACTION |

| Agents (pink) | |
|---|---|
| 1 | PROBLEM_AGENT |
| 2 | THEM_AGENT |
| 3 | US_AGENT |
| 4 | THEM_PRONOUN_AGENT |
| 5 | THEM_PRONOUN_AGENT |
| 6 | US_AGENT |
| 7 | US_AGENT |
| 8 | REF_AGENT |
| 9 | US_AGENT |
| 10 | US_AGENT |
| 11 | US_AGENT |
| 12 | US_AGENT |
| 13 | US_AGENT |
| 14 | US_AGENT |
| 15 | US_AGENT |
| 16 | THEM_PRONOUN_AGENT |
| 17 | US_AGENT |
| 18 | US_AGENT |
| 19 | US_AGENT |

Figure B.1: Agent and Action Types for the Text on p. 300

# B.8. Automatic Annotation (Naive Bayes)

## Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

### Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

### Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

### Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fiddicth (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

# B.9. Automatic Annotation (N-Gram)

## Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

### Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

### Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in large enough corpora, the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

### Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

# Appendix C

# Annotation Materials

## C.1. Study I: Guidelines for Human Annotation of Basic Scheme

### Principles of annotation

These guidelines describe a classification scheme for scientific papers which annotates the *ownership* of scientific ideas. Segmentation of ownership identifies segments in the paper where authors describe general statements about the field, other researcher's work and their own work, cf. C.1.

| | |
|---|---|
| BACKGROUND | Generally accepted background knowledge |
| OTHER | Specific other work |
| OWN | Own work: method, results, future work... |

Figure C.1: Overview of annotation scheme

Each of the classes is associated with a colour, and these colours are matched with marker pens. Please use these to mark your judgement on the printout of the papers.

Annotate from the author's perspective and their opinion about what is general, specific and their own claim, even if you might not agree with the portrayal of the situation as presented in the paper.

The unit of annotation is always the whole sentence. Annotation is mutually exclusive and proceeds sentence by sentence: once you have decided to assign a certain class, you can immediately go to the next sentence, as a sentence cannot have more than one class.

Please annotate all sentences in the abstract, and all sentences in the document except acknowledgement sentences.

## Description of classes

BACKGROUND

BACKGROUND knowledge marks sentences which are presented as uncontroversial in the field. In such sentences, the research context is established. This includes statements of general capacity of the field, general problems, research goals, methodologies and general solutions ("*In recent years, there has been a growing interest in the field of X in the subject of Y*"). The most prototypical use of BACKGROUND is in the beginning of the paper.

Examples for general problems:

- *One of the difficult problems in machine translation from Japanese to English or other European languages is the treatment of articles and numbers.*
- *Complications arise in spelling rule application from the fact that, at compile time, neither the lexical nor the surface form of the root, nor even its length, is known.*
- *Collocations present specific problems in translation, both in human and automatic contexts.*

Examples for generally accepted/old solutions or claims:

- *Tagging by means of a Hidden Markov Model (HMM) is widely recognised as an effective technique for assigning parts of speech to a corpus in a robust and efficient manner.*
- *Current research in lexical aquisition is eminently knowledge-based.*
- *Literature in psychology has amply demonstrated that children do not acquire [...]*

In linguistics papers, mark the description of the linguistic phenomena being covered as BACKGROUND. This includes example sentences. In contrast, the *analysis* of the phenomena are typically either own or other work.

It may be that there is a BACKGROUND segment somewhere in the middle of the paper. It may then not be easy to decide if it is BACKGROUND or OWN. Use the following test: if you think that this segment could have been used as an introductory text at the beginning of the paper, and if it does not contain material that is individualized to the authors themselves, then it should be marked as BACKGROUND.

References to "pioneers" in the field are also BACKGROUND material—sentences which describe other work in an introductory way without any criticism. These are usually older references.

Sometimes there is no BACKGROUND segment, namely if the authors start directly by describing one specific individualized approach.

OTHER

The difference between BACKGROUND and OTHER is only in degree of *specificity*.

OTHER are descriptions of other work which is described *specifically* enough to contrast the own work to it, to criticize it or to mention that it provides support for own idea. For some work to be considered specific other work, it must be clearly attributable to some other researchers, otherwise it might be too general to count as specific other work. Often such segments are started by markers of specific work, citations:

- *<REF> argues that children don't acquire grammar frames until they have a lexicon [...]*
- *<REF> 's solution solves the problem of data-sparseness.*
- *<REF> 's formalism allows the treatment of coordinated structures.*
- *The bilingual dual-coding theory <REF> partially answers the above questions.*
- *<REF> introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse.*

Named solutions can also count as specificity markers for other work:

- *Similarity-based models suggest an appealing approach for dealing with data sparseness.*

The distinction between BACKGROUND and OTHER might be difficult to make. Stop marking as BACKGROUND when you reach a point where ideas, solutions, or tasks are clearly being individualized, i.e. attributed to researchers in such a way that they can get criticized. Often the breaking point looks like this: "*<General problem description> Recently, some researchers have tried to tackle this by doing <More specific description with references>*" In that case, the border is before *"Recently"*.

When authors give specific information about research, but express no stance towards that work, particularly if it happens in the beginning, they seem to imply the statements are generally accepted in the field. You might in this case decide to mark it as BACKGROUND.

OWN

Own work in the context of this paper means work presented as performed by the authors *in the given paper*, i.e. as new research. This includes a description of the own solution, results, discussion, limitations and future work.

*Previous* own research, i.e. research done by the authors before and published elsewhere, does *not* count as own work. Sometimes the fact that previous work is discussed is specifically marked ("*we have previously*"), sometimes it can only be inferred because there is a reference indicating the author's name. Check the reference list to make sure that the string "*et al.*" in a citation (cited paper) does not "hide" one of the authors of the current paper. Unfortunately, authors tend to talk about previous own work in much the same way as they do about the current (own) work. This might constitute a problem here. It is your job to decide if certain statements are presented as if they were the contribution of the paper. There is one exception: PhD or MSc theses do not count as published work (otherwise, some entire papers would have to be marked as other work if the paper is a short version of a PhD or MSc thesis).

Sometimes, short descriptions of own work (statements of opinion) appear within sections talking about other work (background or specific). For example, an author might describe a general problem, then individualize the present research by setting the scope within the current work ("*We will here only be interested in VP gapping as opposed to NP gapping*"), then continue describing general specific to VP gapping. These scope declarations should be considered as own work because they talk about the given work/opinions. The grammatical subject in a sentence does not always tell you whether it's own work or not. Sometimes the criticism of other work might look like own opinion ("*However, we are convinced that this is wrong [...]*"). Cases like this should *not* be considered as own work, but as a description of the weaknesses of other work, i.e. it should be marked as OTHER.

In particular, watch out for the first mention of the own work, typically two thirds down in the introduction. Most of the information under the Summary or Conclusion section is normally own work. Sometimes, individual sentences in the conclusion section make direct comparisons with other work, e.g. detailing advantages of the approach. Only mark these as OTHER if the other work is described again, using more than one sentence of description, else mark as OWN.

## When it gets difficult

There are several reasons why the annotation scheme might not work well for a given paper. The writing style in some papers might make it difficult to see the trisection according to intellectual ownership. In some papers however, the scheme's assumptions that research with different ownership (own/other/background) is indeed presented in separate segments in the paper are violated:

- Our model assumes that the author perceives a clear separation between own work and work outside the scope of the paper, and presents work according to that separation. However, if the paper describes some minute detail of a previous, larger work of the author, then this separation might not be given.

- A specialized case of this, and another example of a potential breakdown of the simple model is for evaluation papers, especially where the authors compare several of their own solutions with each other, or if they compare their solution to somebody else's.

- The scheme also assumes that there is *really* some new contribution described in the paper. This is not the case with position or review articles.

Please keep a note of all difficulties that you encounter with determining individualized segments, and write down your reasons for finding it difficult (i.e. in which way the given paper made it hard for our model to describe what was going on).

# Example annotation

## A Robust Parser Based on Syntactic Information

Kong Joo Lee Cheol Jung Kweon Jungyun Seo Gil Chang Kim

Abstract

An extragrammatical sentence is what a normal parser fails to analyze. It is important to recover it using only syntactic information although results of recovery are better if semantic factors are considered. A general algorithm for least-errors recognition, which is based only on syntactic information, was proposed by G. Lyon to deal with the extragrammaticality. We extend this algorithm to recover extragrammatical sentence into grammatical one in running text. Our robust parser with recovery mechanism - extended general algorithm for least errors recognition - can be easily scaled up and modified because it utilize only syntactic information. To upgrade this robust parser we proposed heuristics through the analysis of the Penn treebank corpus. The experimental result shows 68% ~ 78% accuracy in error recovery.

### 1 Introduction

Extragrammatical sentences include patently ungrammatical constructions as well as utterances that may be grammtically acceptable but are beyond the syntactic coverage of the parser, and any other difficult ones that are encountered in parsing (Carbonell and Hayes, 1983)

I am sure this is what he means.
This, I am sure, what he means.

The progress of machine does not stop even a day.
Not even a day does the progress of machine stop.

Above examples show that people are used to write same meaningful sentence differently. In addition, people are prone to mistakes in writing sentences. So, the bulk of written sentences are open to the extragrammaticality. In the Penn treebank tree-tagged corpus (Marcus, 1991), for instance, about 80 percents of the rules are concerned with peculiar sentences which include inversive, elliptic, paranthetic, or emphatic phrases. For example, we can drive a rule VP -> vb NP comma rb comma PP from the following sentence.

The same jealousy can breed confusion, however, in the absence of any authorization bill this year.

A robust parser is one that can analyze these extragrammatical sentences without failure. However, if we try to preserve robustness by adding such rules whenever we encounter an extragrammatical sentence, the rulebase will grow up rapidly, and thus processing and maintaining the excessive number of rules will become inefficient and impractical. Therefore, extragrammatical sentences should be handled by some recovery mechanism(s) rather than by a set of additional rules.

Many researchers have attempted several techniques to deal with extragrammatical sentences such as Augmentel Transition Networks (ATN) (Kwasny and Sondheimer, 1981), network-based semantic grammar (Hendrix, 1977), partial pattern matching (Hayes and Mouradian, 1981), conceptual case frame (Schank et al, 1980), and multiple cooperative methods (Hayes and Carbonell, 1981). Above mentioned techniques take into account various semantic factors depending on specific domains on question in recovering extragrammatical sentences. Whereas they can provide even better solutions intrinsically, they are usually ad-hoc and are lack of extensibility. Therefore, it is important to recover extragrammatical sentences using syntactic factors only, which are independent of any particular system and any particular domain.

Mellish (Mellish, 1989) introduced some chart-based techniques using only syntactic information for extragrammatical sentences. This technique has an advantage that there is no repeating work for the chart to prevent the parser from generating the same edge as the previously existed edge. Also, because the recovery process runs when a normal parser terminates unsuccessfully, the performance of the normal parser does not decrease in case of handling grammatical sentences. However, his experiment was not based on the errors in running texts but on artificial ones which were randomly generated by human. Moreover, only one word error was considered though several word errors can occur simultaneously in the running text.

A general algorithm for least-errors recognition (Lyon, 1974) proposed by G.Lyon, is to find out the least number of errors necessary to sucessful parsing and recover them. Because this algorithm is also syntactically oriented and based on a chart, it has the same advantage as Mellish's parser. When the original parsing algorithm terminates unsuccessfully, the algorithm begins to assume errors of insertion, deletion and mutation of a word. For any input, this algorithm can generate the resultant parse tree. At the cost of the complete robustness, however, this algorithm degrades the efficiency of parsing, and generates many intermediate edges.

In this paper, we present a robust parser with a recovery mechanism. We extend the general algorithm for least-error recognition to adopt it as the recovery mechanism in our robust parser. Because our robust parser handle extragrammatical sentences with this syntactic information oriented recovery mechanism, it can be independent of a particular system or particular domain. Also, we present the heuristics to reduce the number of edges so that we can upgrade the performance of our parser.

This paper is organized as follows: We first review a general algorithm for least-errors recognition. Then we present the extension of this algorithm, and the heuristics adopted by the robust parser. Next, we describe the implementation of the system and the result of the experiment of parsing real sentences. Finally, we make conclusion with future direction.

### 4 Conclusion

In this paper, we have presented the robust parser with the extended least-errors recognition algorithm as the recovery mechanism. This robust parser can easily be scaled up and applied to various domains because this parser depends only on syntactic factors. To enhance the performance of the robust parser for extragrammatical sentences, we proposed several heuristics. The heuristics assign the error values to each error-hypothesis edge, and edges which has less error are processed first. So, not all the generated edges are processed by the robust parser, but the most plausible parse trees can be generated first. The accuracy of the recovery of our robust parser is about 68% ~ 77 %. Hence, this parser is suitable for systems in real application areas.

Our short term goal is to propose an automatic method that can learn parameter values of heuristics by analyzing the corpus. We expect that automatically learned values of parameters can upgrade the performance of our parser.

Acknowledgement

This work was supported (in part) by Korea Science and Engineering Foundation (KOSEF) through Centre for Artificial Intelligence Research (CAIR), the Engineering Research Centre (ERC) of Excellence Program.

### References

[Black, 1991] E. Black et. al. A procedure for quantitatively comparing the syntactic coverage of English Grammars. Proceedings of Fourth DARPA Speech and Natural Language Workshop, 1991.

[Carbonell and Hayes, 1983] J. G. Carbonell and P. J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. American Journal of Computational Linguistics, vol. 9, no 3-4, 1983.

[Hayes and Carbonell, 1981] P. Hayes and J. Carbonell. Multi-strategy Construction-Specific Parsing for Flexible Data Base Query Update. Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

[Hayes and Mouradian, 1981] P. J. Hayes and G. V. Mouradian. Flexible Parsing. American Journal of Computational Linguistics, vol. 7, no. 4, 1981.

[Hendrix, 1977] G. Hendrix. Human Engineering for Applied Natural Language Processing. Proceedings of the 5th International Joint Conference on Artificial Intelligence, 1977.

[Kwasny and Sondheimer, 1981] S. Kwasny and N. Sondheimer. Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. American Journal of Computational Linguistics, vol. 7, no. 2, 1981.

[Lyon, 1974] G. Lyon. Syntax-Directed Least-Errors Analysis for Context-Free languages. Communications of the ACM, vol. 17, no. 1, 1974.

[Marcus, 1991] M. P. Marcus. Building very large natural language corpora: The Penn Treebank, 1991.

[Mellish, 1989] C. S. Mellish. Some Chart-Based Techniques for Parsing Ill-Formed Input. Association for Computational Linguistics, 1989.

[Schank et al, 1980] R.C. Schank et al. An Integrated Understander. American Journal of Computational Linguistics, vol 6, no.1, 1980

GENERAL

OTHER

OWN

# Splitting the reference time: Temporal Anaphora and Quantification in DRT

Rani Nelken
Nissim Francez

Abstract

This paper presents an analysis of temporal anaphora in sentences which contain quantification over events, within the framework of Discourse Representation Theory. The analysis in (Partree, 1984) of quantified sentences, introduced by a temporal connective, gives the wrong truth-conditions when the temporal connective in the subordinate clause is before or after. This problem has been previously analyzed in (de Swart, 1991) as an instance of the proportion problem, and given a solution from a Generalized Quanitifier approach. By using a careful distinction between the different notions of reference time, based on (Kamp and Reyle, 1993), we propose a solution to this problem, within the framework of DRT. We show some applications of this solution to additional temporal anaphora phenomena in quantified sentences.

## 1 Introduction

The analysis of temporal expressions in natural language discourse provides a challenge for contemporary semantics theories. (Partree, 1973) introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse. In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities. Such interaction, while interesting in its own right, is also a good test-bed for theories of the semantic interpretation of temporal expressions. We discuss cases such as:

(1) Before John makes a phone call, he always lights up a cigarette (Partree, 1984).

(2) Often, when Anne came home late, Paul had already prepared dinner. (de Swart, 1991)

(3) When he came home, he always switched on the TV. He took a beer and sat down in his armchair to forget the day. (de Swart, 1991)

(4) When John is at the beach, he always squints when the sun is shining. (de Swart, 1991)

The analysis of sentences such as (1) in (Partree, 1984), within the framework of Discourse Representation Theory (DRT) (Kamp, 1981) gives the wrong thruth-conditions, when the temporal connective in the sentence is before or after. In DRT, such sentences trigger box-splitting with the eventuality of the subordinate clause and an updated refernece time in the antecedent box, and the eventuality of the main clause in the consequent box, causing undesirable universal quantification over the reference time.

This problem is analyzed in (de Swart, 1991) as an instance of the proportion problem and given a solution from a Generalized Quanifier approach. We were led to seek a solution for this problem within DRT, because of DRT's advantages as a general theory of discourse, and its choice as the underlying formalism in another research project of ours, which deals with sentences such as 1-4, in the context of natural language specifications of computerized systems. In this paper, we propose such a solution based on a careful distinction between different roles of Reichenbach's reference time (Reichenbach, 1947), adapted from (Kamp and Reyle, 1993). Figure 1 shows a 'minimal pair' of DRS's for sentence 1, one according to Partee's (1984) analysis and one according to ours.

## 2 Background

An analysis of the mechanism of temporal anaphoric reference hinges upon an understanding of the ontological and logical foundations of temporal reference.

## C.2. Study II: Guidelines for Human Annotation of Full Scheme

These guidelines describe a classification scheme for scientific papers for ownership of ideas, relation to other work and internal paper structure. The classification scheme is displayed in Figure C.2.

Each of the classes is associated with a colour, and these colours are matched with marker pens. Please use these to mark your judgement on the printout of the papers.

| | |
|---|---|
| BACKGROUND | Generally accepted background knowledge |
| OTHER | Specific other work |
| OWN | Own work: method, results, future work... |
| AIM | Specific research goal |
| TEXTUAL | Textual section structure |
| CONTRAST | Contrast, comparison, weakness of other solution |
| BASIS | Other work provides basis for own work |

Figure C.2: Overview of annotation scheme

### Annotation procedure

#### Before annotation

Skim-read the paper before annotation. This is important, as in some papers, the interpretation of certain sentences in the context of the overall argumentation only becomes apparent after one has an overview of the whole paper. Don't try to understand the solution in detail—you can jump over the parts of the paper where you think the own solution is described in details. Rather try to understand the structure of the scientific argumentation. Concentrate on those parts of the paper where the connection to the subject field and the connection to other work is described. In particular, skim-read the abstract, the introduction, the conclusions (if it is summary-style), and sections reviewing other research (often after introduction or before conclusions; they could be marked sections with headlines like “Relation to other work”, “Prior research”, “X in the literature” etc.).

### Annotation procedure

Annotation proceeds sentence by sentence, and is mutually exclusive: Each sentence can have only one category. The main decision procedure is given in Figure C.3. For each sentence, the following questions have to be answered.

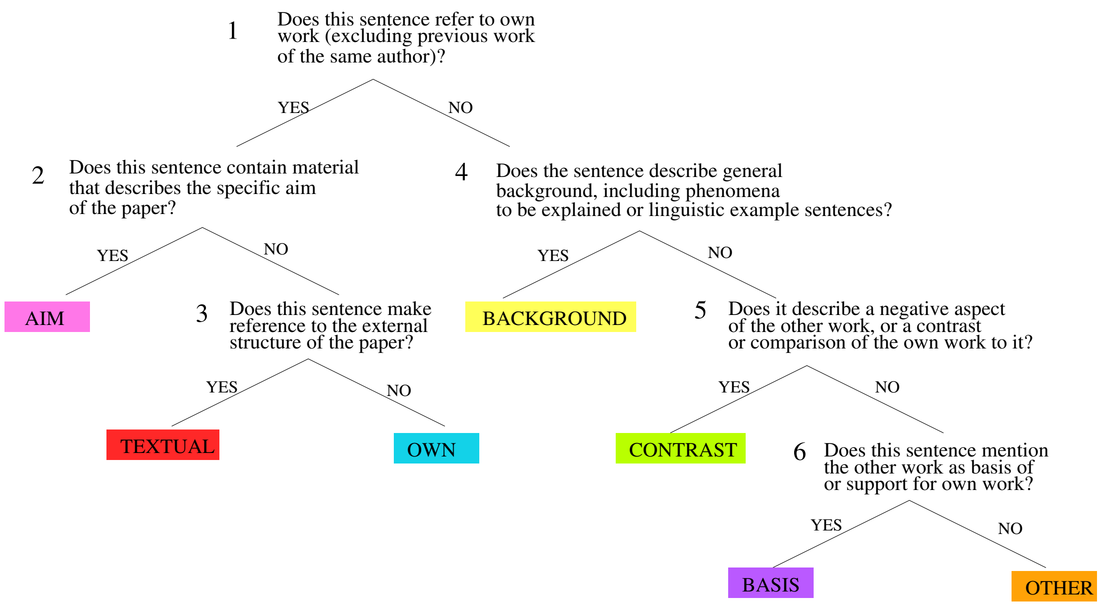



Figure C.3: Decision process

Therefore, if there is a conflict, the “higher” classes in the decision tree (the ones that you reach first) will win over the “lower” classes. These guidelines will give details about the questions.

When interpreting the role of a sentence, you should treat the sentence in the way in which you think the *author* intended it in their argumentation. Context and location of a sentence are important.

- **Question 1: Does this sentence talk about own work?**
  If your answer is ’yes’, proceed to Question 2.
  If your answer is ’no’, proceed to Question 4.

- **Question 2: Does it contain a goal statement?**
  If your answer is 'yes', assign class AIM and move to next sentence.
  If your answer is 'no', proceed to Question 3.

- **Question 3: Does it contain a textual overview?**
  If your answer is 'yes', assign tag TEXTUAL and move to the next sentence.
  If your answer is 'no', assign tag OWN and move to the next sentence.

- **Question 4: Does it describe background?**
  If your answer is 'yes', assign tag BACKGROUND and move to the next sentence.
  If your answer is 'no', proceed to Question 6.

- **Question 5: Is the other work described in a contrastive way?**
  If your answer is 'yes', assign tag CONTRAST and move to next sentence.
  If your answer is 'no', proceed to Question 5.

- **Question 6: Is the own work based on other work?**
  If your answer is 'yes', assign tag BASIS.
  If your answer is 'no', assign tag OTHER.

You can mark consecutive sentences with the same category if they *together* fulfill the criteria of the category. E.g. you could mark two sentences as AIM if they together describe the specific goal of a paper well. If you cannot assign a category, please mark the sentence and take a note describing the difficulties.

As soon as you have reached a leaf, assign the corresponding category to the sentence. Please annotate all sentences in the abstract, and all sentences in the document except acknowledgement sentences. Also mark (linguistic) example sentences.

### After annotation

Check a few things, and rectify your annotation if necessary:

- There must be at least one AIM sentence. If this is not the case, reclassify some other candidate sentences, until you have found at least one sentence that represents the specific aim of the given paper.

- There must not be more than 5 AIM sentences per paper. The only exception is if each of them is a straight hit, i.e. they are indisputably goal statements, particularly if the sentences are paraphrases of each other.

  If you have to eliminate AIM sentences, do the following:

  - Prefer explicit AIM statements (prefer 'direct' goal statements and 'functionality-provided' to 'solved' and other types).
  - Prefer AIM sentences towards the periphery (e.g. at the beginning of summarizing conclusions), and in the border area with OTHER or Background segments;
  - If all fails, pick the ones you think are most relevant in the context of distinguishing this piece of research from others.

## The questions

### Question 1: Does this sentence talk about own work?

Own work in the context of this paper means work presented as performed by the authors *in the given paper*, i.e. as new research.

Description of own work should make up a large part of the paper—it includes descriptions of the own solution, method, results, discussion, limitations and future work.

*Previous* own research, i.e. research done by the authors before and published elsewhere, does *not* count as own work. Sometimes the fact that previous work is discussed is specifically marked ("*we have previously*"), sometimes it can only be inferred because there is a reference indicating the author's name. Check the reference list to make sure that the string "*et al.*" in a citation (cited paper) does not "hide" one of the authors of the current paper. Unfortunately, authors tend to talk about previous own work in much the same way as they do about the current (own) work. This might constitute a problem here. It is your job to decide if certain statements are presented as if they were the contribution of the paper. There is one exception: PhD or MSc theses do not count as published work (otherwise, some entire papers would have to be marked as other work if the paper is a short version of a PhD or MSc thesis). In that case, the sentence first citing the thesis is to be marked as BASIS. In all other contexts, reference to the thesis/research is to be considered as own.

Sometimes, short descriptions of own work (statements of opinion) appear within sections talking about other work (background or specific). For example, an author might describe a general problem, then individualize the present research by setting the scope within the current work ("*We will here only be interested in VP gapping as opposed to NP gapping*"), then continue describing general specific to VP gapping. These scope declarations should be considered as own work because they talk about the given work/opinions. The grammatical subject in a sentence does not always tell you whether it's own work or not. Sometimes the criticism of other work might look like own opinion ("*However, we are convinced that this is wrong [...]*"). Cases like this should *not* be considered as own work, but as weaknesses of other work, i.e. OTHER.

In particular, watch out for the first mention of the own work, typically two thirds down in the introduction. Most of the information under the Summary or Conclusion section is normally own work. Sometimes, individual sentences in the conclusion section make direct comparisons with other work, e.g. detailing advantages of the approach. Only mark these as OTHER if the other work is described again, using more than one sentence of description, else mark as OWN.

**Question 2: Does this sentence contain a goal statement?**

Two kinds of sentences count as goal statements:

- Goal statements (i.e. description of research goal)
- Scope statement (i.e. delimitation of research goal: what the goal is not)

If the sentence describes a general goal in the field, e.g. *"machine translation"*, it should not be marked as AIM. AIM sentences describe *particular* goals of the paper. There are different ways of expressing the particular goal of the paper.

A prime location of AIM sentences is around the first 2/3 of the introduction, when the authors are mentioned for the first time.

**Direct aim/goal description:**

- *Our aim in this paper is to [...]*
- *We, in contrast, aim at defining categories that help us [...]*

Also descriptions of phenomena plus the statement that current work tries to explain them, e.g.:

- *We aim to find a method of inducing grammar rules.*
- *Our goal, however, is to develop a mechanism for [...]*
- *We will introduce PHENOMENON X that we seek to explain*
- *I show how grammar rules can be induced.*

**Functionality provided:** Another way of expressing the research goal is to say that one has accomplished doing a certain task.

- *This paper gives a syntactic-head-driven generation algorithm which includes a well-defined treatment of moved constituents.*
- *We have presented an analysis of the data sparseness problem*
- *I have presented an analysis of PHENOMENON X*
- *We have presented an analysis of why children cannot [...]* (PHENOMENON)

**Hypothesis:** In experimental papers the goal might be expressed as a hypothesis:

- *The hypothesis investigated in this paper is that children can acquire [...]*

**Goal as focus:** The declaration of a research interest can count as an $\boxed{\text{AIM}}$:

- *This paper focuses on inducing grammar rules.*
- *This paper concerns the formal definitions underlying synchronous tree-adjoining grammars.*
- *In this paper, we focus on the application of the developed techniques in the context of the comparatively neglected area of HPSG generation.*
- *This paper will focus on [...] our analysis of narrative progression, rhetorical structure, perfects and temporal expressions.*

**Solutionhood:** Sometimes a sentence states that the own solution works, i.e. solves a particular research task. Such sentences can under certain circumstances be AIM s, but they are AIM s of a lower quality. You must be sure that the announcement of the successful problem-solving process is indeed important enough to cover the goal of the whole paper, and you must be sure that the sentence refers to the *highest* level of problem solving. If it talks about a *sub*problem, don't consider the sentence an AIM . Often such statements are dressed as a claim.

Examples:

- *[we present an analysis] which automatically gives the right results for quantifier scope ambiguities and interactions with bound anaphora.*
- *In this paper we presented a new model that implements the similarity-based approach to provide estimates for the conditional probabilities of unseen word cooccurrences*
- *Our technique segments continuous speech into words using only distributional and phonotactic information*
- *The Spoken Language Translator (SLT) is a prototype system that translates air travel (ATIS) queries from spoken English to spoken Swedish and to French.*

**Definition of a desired property or as necessity:** The goal can be given by describing a hypothetical, desired mechanism or a desired outcome. This is not a typical way to describe the paper's AIM , but the context can still make this the "best AIM around".

Examples:

- *A robust Natural Language Processing (NLP) system must be able to process sentences that contain words unknown to its lexicon.*
- *The importance of a method for SPECIFIC-TASK grows as the coverage of [...] improves.*
- *and I demonstrate the importance of having a Y tool which allows for X.*

**Advantage of a solution:** Sometimes the description of an advantage of a solution can provide an acceptable AIM :

- *Our method yields polynomial complexity in an elegant way.*
- *Our method avoids problems of non-determinacy.*

- *First, it is in certain respects simpler, in that it requires no postulation of otherwise unmotivated ambiguities in the source clause.*
- *The traditional problems of training times do not arise.*

**Scope statement:** These sentences define the goal as *part* of previous goal, e.g. *"here we will look only at relative pronouns"*, excluding some other, similar goals.

**Indirect aim/goal description:** In some cases, if you find nothing better, you can also look for more indirect ways of expressing what the goal might have been.

- *In this paper we address two issues relating to the application of preference functions.*
- *[...] and make a specific proposal concerning the interface between these and the syntactic and semantic representations they utilize.*
- *In addition, we have taken a few steps towards determining the relative importance of different factors to the successful operation of discourse modules.*

**Question 3: Does this sentence contain a textual overview?**

All statements whose primary function it is to give us an overview of the section structure (*"in the next section we will [...]"*). Several such sentences often occur at the end of the introduction.

Mark also backward looking pointers at the beginning of a section (first sentence) (*"In the previous section we have implemented a model"*) or before the end of the section ("*in the next section, we will turn our attention to [...] "*. Some authors give an overview of the section at the beginning of the section ("*in this section I will [dots]*"), or summarize after each section ("*in this section I have [dots]*" or "*this concludes my discussion of X"*.

Caveat: Sentences referring to figures or tables are not meant here (*"figure 3 shows [...]*")!

Sentences summing up main conclusions from *previous* sections are also not meant here:

- *"In chapter 3, we have seen that children cannot reliably form generalizations about [...]".*

**Question 4: Does this sentence describe background?**

BACKGROUND knowledge marks sentences which are presented as uncontroversial in the field. In such sentences, the research context is established. This includes statements of general capacity of the field, general problems, research goals, methodologies and general solutions ("*In recent years, there has been a growing interest in the field of X in the subject of Y*"). The most prototypical use of BACKGROUND is in the beginning of the paper.

Examples for general problems:

- *One of the difficult problems in machine translation from Japanese to English or other European languages is the treatment of articles and numbers.*
- *Complications arise in spelling rule application from the fact that, at compile time, neither the lexical nor the surface form of the root, nor even its length, is known.*
- *Collocations present specific problems in translation, both in human and automatic contexts.*

Examples for generally accepted/old solutions or claims:

- *Tagging by means of a Hidden Markov Model (HMM) is widely recognised as an effective technique for assigning parts of speech to a corpus in a robust and efficient manner.*
- *Current research in lexical aquisition is eminently knowledge-based.*
- *Literature in psychology has amply demonstrated that children do not acquire [...]*

In linguistics papers, mark the description of the linguistic phenomena being covered as BACKGROUND. This includes example sentences. In contrast, the *analysis* of the phenomena are typically either own or other work.

It may be that there is a BACKGROUND segment somewhere in the middle of the paper. It may then not be easy to decide if it is BACKGROUND or OWN. Use the following test: if you think that this segment could have been used as an introductory text at the beginning of the paper, and if it does not contain material that is individualized to the authors themselves, then it should be marked as BACKGROUND.

References to “pioneers” in the field are also $\boxed{\text{BACKGROUND}}$ material—sentences which describe other work in an introductory way without any criticism. These are usually older references.

Sometimes there is no $\boxed{\text{BACKGROUND}}$ segment, namely if the authors start directly by describing one specific individualized approach.

The difference between $\boxed{\text{BACKGROUND}}$ and $\boxed{\text{OTHER}}$ is only in degree of *specificity*.

$\boxed{\text{OTHER}}$ are descriptions of other work which is described *specifically* enough to contrast the own work to it, to criticize it or to mention that it provides support for own idea. For some work to be considered specific other work, it must be clearly attributable to some other researchers, otherwise it might be too general to count as specific other work. Often such segments are started by markers of specific work, citations:

- *<REF> argues that children don't acquire grammar frames until they have a lexicon [...]*
- *<REF> 's solution solves the problem of data-sparseness.*
- *<REF> 's formalism allows the treatment of coordinated structures.*
- *The bilingual dual-coding theory <REF> partially answers the above questions.*
- *<REF> introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse.*

Named solutions can also count as specificity markers for other work:

- *Similarity-based models suggest an appealing approach for dealing with data sparseness.*

The distinction between $\boxed{\text{BACKGROUND}}$ and $\boxed{\text{OTHER}}$ might be difficult to make. Stop marking as $\boxed{\text{BACKGROUND}}$ when you reach a point where ideas, solutions, or tasks are clearly being individualized, i.e. attributed to researchers in such a way that they can get criticized. Often the breaking point looks like this: “<*General problem description*> *Recently, some researchers have tried to tackle this by doing* <*More specific description with references*>” In that case, the border is before *“Recently”*.

When authors give specific information about research, but express no stance towards that work, particularly if it happens in the beginning, they seem to imply the statements are generally accepted in the field. You might in this case decide to mark it as BACKGROUND.

**Question 5: Is the other work described in a contrastive way?**

These sentences make one type of connection between specific other work and own work. Comparative sentences might occur within segments describing other work or own work (e.g. in conclusions).

Mark sentences which contain mentions of:

- Weaknesses of other people's solutions
- The absence of a solution for a given problem
- Difference in approach/solution
- Superiority of own solution
- Statements of direct comparisons with other work or between several other approaches (these appear mostly in evaluation papers)
- Incompatibility between own and other claims or results

**Weaknesses of other solutions:**

- *<REF>'s solution is problematic for several reasons.*
- *The results suggest that a completely unconstrained initial model does not produce good quality results.*
- *Here, we will produce experimental evidence suggesting that this simple model leads to serious overestimates of system error rates.*
- *The analysis of sentences such as <CREF> in <REF>, within the framework of Discourse Representation Theory (DRT) <REF> gives the wrong truth-conditions, when the temporal connective in the sentence is "before" or "after".*
- *A limiting factor of this method is the potentially large number of distinct parse trees.*

**Absence of a solution:**

- *While we know of previous work which associates scores with feature structures <REF> we are not aware of any previous treatment which makes explicit the link to classical probability theory.*
- *First, although much work has been done on how agents request clarifications, or respond to such requests, little attention has been paid to the collaborative aspects of clarification discourse.*

**Difference in approach/solution:**

- *In contrast to standard approaches, we use a statistical model.*
- *In this paper, we propose an alternative approach in which a performance-oriented (behaviour-based) perspective is taken instead of a competence-oriented (knowledge-based) one.*
- *Namely, since we use semantic/pragmatic roles instead of grammatical roles in constraints [...]*

**Superiority of own solution:**

- *Our model outperforms simple pattern-matching models by 25%.*
- *Our results indicate that our full integrated heuristic scheme for selecting the best parse out-performs the simple heuristic [...]*
- *We have also argued that an architecture that uses obligations provides a much simpler implementation than the strong plan-based approaches.*

**Direct comparisons with other work:**

- *In this paper, we will compare two tagging algorithms, one based on classifying word types, and one based on classifying words-plus-context.*
- *[...] and a comparison with manual scaling in section <CREF>.*
- *The performance of both implementations is evaluated and compared on a range of artificial and real data.*

**Incompatibility between own and other claims or results:**

- *This result challenges the claims of recent discourse theories (<REF>, <REF>) which argue for a the close relation between cue words and discourse structure.*
- *It is implausible that children learn grammar on the fly.*

There is a conflict between AIM and CONTRAST when goals are introduced contrastively, as in the following examples. These sentences would normally be tagged AIM, unless there are too many better AIM sentences around.

- *Until now, research has focused on demonstrations of infants' sensitivity to various sources; we have begun to provide quantitative measures of the usefulness of those sources.*
- *However our objective is not to propose a faster algorithm, but is to show the possibility of distributed processing of natural languages.*
- *This article proposes a method for automatically finding the appropriate tree-cutting criteria in the EBG scheme, rather than having to hand-code them.*

If the sentence expresses no sentential content other than the fact that there is a contrast ("*however, our approach is quite different*") mark this sentence only as CONTRAST if you don't find a better one.

If authors compare their own work contrastively to somebody else's (e.g. a linguistic analysis) to explain in which aspects their own work is superior, you might be undecided as to whether to mark it as CONTRAST or OWN (or even AIM, in some cases!). Assign AIM only if the authors specifically say that they did something differently in order to achieve a (different?) goal. Assign CONTRAST if you believe that the main function of the sentence is to mention a negative aspect of the other work. Assign OWN if the focus is on their own work rather than on the other work.

**Question 6: Is the own work based on other work?**

There are 5 different classes of how work could be based or positively related:

- Direct Based
- Adaptation

- Consistency
- Similarity
- Quality

Consistency, Similarity and Quality cases should be marked only if the approaches are important to the paper, i.e. if some more discussion about that work is given in the paper.

**Direct Based:** It is explicitly stated that the own solution builds on another solution (intellectual ancestry).

- *We base our model on <REF>'s backup model.*
- *Our approach is in the spirit of <REF> 's approach*
- *We choose to use Link Grammar <REF>*

The last example describes a BASIS describing intellectual ancestry with more than one other approach.

**Adaptation:** The authors have adapted a solution, contributed by somebody else. As the solution was not initially invented for the current research task, and needs to be adapted.

- *The main aim is to show how existing text planning techniques can be adapted for this particular application.*
- *We extend the model for doing X by allowing it to do Y, too.*
- *We have suggested some ways in which LFs can be enriched with lexical semantic information to improve translation quality.*
- *This model draws upon <REF>, but adapts it to the collaborative situation.*
- *In our work, we have taken <REF>'s descriptive model and recast it into a computational one [...]*

**Consistency:** Statements about consistency with another theoretical framework or other people's results can be BASIS, even if the own solution is not directly based on it:

- *Our account [...] fits within a general framework for [...]*

**Similarity:** Statements about similarities between the own and other approaches can be a $\boxed{\text{BASIS}}$, if these similarities are not "cancelled" later by mentioning a contrasting property.

- *The analysis presented here has strong similarities to analyses of the same phenomena discussed by <REF> and <REF>.*
- *The method, which is related to that of <REF>,*
- *In this section we define a grammar similar to <REF>'s first grammar.*

**Quality of other approach:** If you think that an approach provides a basis, and is important enough to be marked up as a $\boxed{\text{BASIS}}$, but you can find no explicit sentence expressing it, you can mark up statements about the quality of the approach.

- *We discuss the advantages of <REF>'s model.*
- *[...] the success of an abstract model such as <REF>'s [...]*
- *[...] thus demonstrating the computational feasibility of their work and its compatibility with current practices in artificial intelligence.*
- *Earley deduction is a very attractive framework for natural language processing because it has the following properties and applications.*

# Example annotation

## A Robust Parser Based on Syntactic Information

Kong Joo Lee  Cheol Jung Kweon  Jungyun Seo  Gil Chang Kim

### Abstract

An extragrammatical sentence is what a normal parser fails to analyze. It is important to recover it using only syntactic information although results of recovery are better if semantic factors are considered. A general algorithm for least-errors recognition, which is based only on syntactic information, was proposed by G. Lyon to deal with the extragrammaticality. We extend this algorithm to recover extragrammatical sentence into grammatical one in running text. Our robust parser with recovery mechanism - extended general algorithm for least errors recognition - can be easily scaled up and modified because it utilize only syntactic information. To upgrade this robust parser we proposed heuristics through the analysis of the Penn treebank corpus. The experimental result shows 68% ~ 78% accuracy in error recovery.

### 1 Introduction

Extragrammatical sentences include patently ungrammatical constructions as well as utterances that may be grammtically acceptable but are beyond the syntactic coverage of the parser, and any other difficult ones that are encountered in parsing (Carbonell and Hayes, 1983)

I am sure this is what he means.
This, I am sure, what he means.

The progress of machine does not stop even a day.
Not even a day does the progress of machine stop.

Above examples show that people are used to write same meaningful sentence differently. In addition, people are prone to mistakes in writing sentences. So, the bulk of written sentences are open to the extragrammaticality. In the Penn treebank tree-tagged corpus (Marcus, 1991), for instance, about 80 percents of the rules are concerned with peculiar sentences which include inversive, elliptic, paranthetic, or emphatic phrases. For example, we can drive a rule VP -> vb NP comma rb comma PP from the following sentence.

The same jealousy can breed confusion, however, in the absence of any authorization bill this year.

A robust parser is one that can analyze these extragrammatical sentences without failure. However, if we try to preserve robustness by adding such rules whenever we encounter an extragrammatical sentence, the rulebase will grow up rapidly, and thus processing and maintaining the excessive number of rules will become inefficient and impractical. Therefore, extragrammatical sentences should be handled by some recovery mechanism(s) rather than by a set of additional rules.

Many researchers have attempted several techniques to deal with extragrammatical sentences such as Augmented Transition Networks (ATN) (Kwasny and Sondheimer, 1981), network-based semantic grammar (Hendrix, 1977), partial pattern matching (Hayes and Mouradian, 1981), conceptual case frame (Schank et al, 1980), and multiple cooperative methods (Hayes and Carbonell, 1981). Above mentioned techniques take into account various semantic factors depending on specific domains on question in recovering extragrammatical sentences. Whereas they can provide even better solutions intrinsically, they are usually ad-hoc and are lack of extensibility. Therefore, it is important to recover extragrammatical sentences using syntactic factors only, which are independent of any particular system and any particular domain.

Mellish (Mellish, 1989) introduced some chart-based techniques using only syntactic information for extragrammatical sentences. This technique has an advantage that there is no repeating work for the chart to prevent the parser from generating the same edge as the previously existed edge. Also, because the recovery process runs when a normal parser terminates unsuccessfully, the performance of the normal parser does not decrease in case of handling grammatical sentences. However, his experiment was not based on the errors in running texts but on artificial ones which were randomly generated by human. Moreover, only one word error was considered though several word errors can occur simultaneously in the running text.

A general algorithm for least-errors recognition (Lyon, 1974) proposed by G.Lyon, is to find out the least number of errors necessary to sucessful parsing and recover them. Because this algorithm is also syntactically oriented and based on a chart, it has the same advantage as Mellish's parser. When the original parsing algorithm terminates un-

insertion, deletion and mutation of a word. For any input, this algorithm can generate the resultant parse tree. At the cost of the complete robustness, however, this algorithm degrades the efficiency of parsing, and generates many intermediate edges.

In this paper, we present a robust parser with a recovery mechanism. We extend the general algorithm for least-error recognition to adopt it as the recovery mechanism in our robust parser. Because our robust parser handle extragrammatical sentences with this syntactic information oriented recovery mechanism, it can be independent of a particular system or particular domain. Also, we present the heuristics to reduce the number of edges so that we can upgrade the performance of our parser.

This paper is organized as follows: We first review a general algorithm for least-errors recognition. Then we present the extension of this algorithm, and the heuristics adopted by the robust parser. Next, we describe the implementation of the system and the result of the experiment of parsing real sentences. Finally, we make conclusion with future direction.

### 4 Conclusion

In this paper, we have presented the robust parser with the extended least-errors recognition algorithm as the recovery mechanism. This robust parser can easily be scaled up and applied to various domains because this parser depends only on syntactic factors. To enhance the performance of the robust parser for extragrammatical sentences, we proposed several heuristics. The heuristics assign the error values to each error-hypothesis edge, and edges which has less error are processed first. So, not all the generated edges are processed by the robust parser, but the most plausible parse trees can be generated first. The accuracy of the recovery of our robust parser is about 68% ~ 77 %. Hence, this parser is suitable for systems in real application areas.

Our short term goal is to propose an automatic method that can learn parameter values of heuristics by analyzing the corpus. We expect that automatically learned values of parameters can upgrade the performance of our parser.

### Acknowledgement

This work was supported (in part) by Korea Science and Engineering Foundation (KOSEF) through Centre for Artificial Intelligence Research (CAIR), the Engineering Research Centre (ERC) of Excellence Program.

### References

[Black, 1991] E. Black et. al. A procedure for quantitatively comparing the syntactic coverage of English Grammars. Proceedings of Fourth DARPA Speech and Natural Language Workshop, 1991.

[Carbonell and Hayes, 1983] J. G. Carbonell and P. J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. American Journal of Computational Linguistics, vol. 9, no 3-4, 1983.

[Hayes and Carbonell, 1981] P. Hayes and J. Carbonell. Multi-strategy Construction-Specific Parsing for Flexible Data Base Query Update. Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981.

[Hayes and Mouradian, 1981] P. J. Hayes and G. V. Mouradian. Flexible Parsing. American Journal of Computational Linguistics, vol. 7, no. 4, 1981.

[Hendrix, 1977] G. Hendrix. Human Engineering for Applied Natural Language Processing. Proceedings of the 5th International Joint Conference on Artificial Intelligence, 1977.

[Kwasny and Sondheimer, 1981] S. Kwasny and N. Sondheimer. Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. American Journal of Computational Linguistics, vol. 7, no. 2, 1981.

[Lyon, 1974] G. Lyon. Syntax-Directed Least-Errors Analysis for Context-Free languages. Communications of the ACM, vol. 17, no. 1, 1974.

[Marcus, 1991] M. P. Marcus. Building very large natural language corpora: The Penn Treebank, 1991.

[Mellish, 1989] C. S. Mellish. Some Chart-Based Techniques for Parsing Ill-Formed Input. Association for Computational Linguistics, 1989.

[Schank et al, 1980] R.C. Schank et al. An Integrated Understander. American Journal of Computational Linguistics, vol 6, no.1, 1980

BKG

OTH

OWN

AIM

CTR

BAS

TXT

# Splitting the reference time: Temporal Anaphora and Quantification in DRT

Rani Nelken
Nissim Francez

Abstract

This paper presents an analysis of temporal anaphora in sentences which contain quantification over events, within the framework of Discourse Representation Theory. The analysis in (Partree, 1984) of quantified sentences, introduced by a temporal connective, gives the wrong truth-conditions when the temporal connective in the subordinate clause is before or after. This problem has been previously analyzed in (de Swart, 1991) as an instance of the proportion problem, and given a solution from a Generalized Quanitifier approach. By using a careful distinction between the different notions of reference time, based on (Kamp and Reyle, 1993), we propose a solution to this problem, within the framework of DRT. We show some applications of this solution to additional temporal anaphora phenomena in quantified sentences.

## 1 Introduction

The analysis of temporal expressions in natural language discourse provides a challenge for contemporary semantics theories. (Partree, 1973) introduced the notion of temporal anaphora, to account for ways in which temporal expressions depend on surrounding elements in the discourse for their semantic contribution to the discourse. In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities. Such interaction, while interesting in its own right, is also a good test-bed for theories of the semantic interpretation of temporal expressions. We discuss cases such as:

(1) Before John makes a phone call, he always lights up a cigarette (Partree, 1984).

(2) Often, when Anne came home late, Paul had already prepared dinner. (de Swart, 1991)

(3) When he came home, he always switched on the TV. He took a beer and sat down in his armchair to forget the day. (de Swart, 1991)

(4) When John is at the beach, he always squints when the sun is shining. (de Swart, 1991)

The analysis of sentences such as (1) in (Partree, 1984), within the framework of Discourse Representation Theory (DRT) (Kamp, 1981) gives the wrong thruth-conditions, when the temporal connective in the sentence is before or after. In DRT, such sentences trigger box-splitting with the eventuality of the subordinate clause and an updated refernece time in the antecedent box, and the eventuality of the main clause in the consequent box, causing undesirable universal quantification over the reference time.

This problem is analyzed in (de Swart, 1991) as an instance of the proportion problem and given a solution from a Generalized Quantifier approach. We were led to seek a solution for this problem within DRT, because of DRT's advantages as a general theory of discourse, and its choice as the underlying formalism in another research project of ours, which deals with sentences such as 1-4, in the context of natural language specifications of computerized systems. In this paper, we propose such a solution based on a careful distinction between different roles of Reichenbach's reference time (Reichenbach, 1947), adapted from (Kamp and Reyle, 1993). Figure 1 shows a 'minimal pair' of DRS's for sentence 1, one according to Partee's (1984) analysis and one according to ours.

## 2 Background

An analysis of the mechanism of temporal anaphoric reference hinges upon an understanding of the ontological and logical foundations of temporal reference.

## C.3. Study III: Short Instructions for Human Annotation

This coding scheme is about the ownership of ideas in scientific papers and about author's stance towards other work. Your intuitions about the structure of this paper will be useful input to help build better tools for information extraction from scientific papers, which in turn will improve automatic bibliographic search.

Read the complete paper first to get a sense of what it is about. You do not have to understand the details of the paper. Then, working from the beginning, mark each

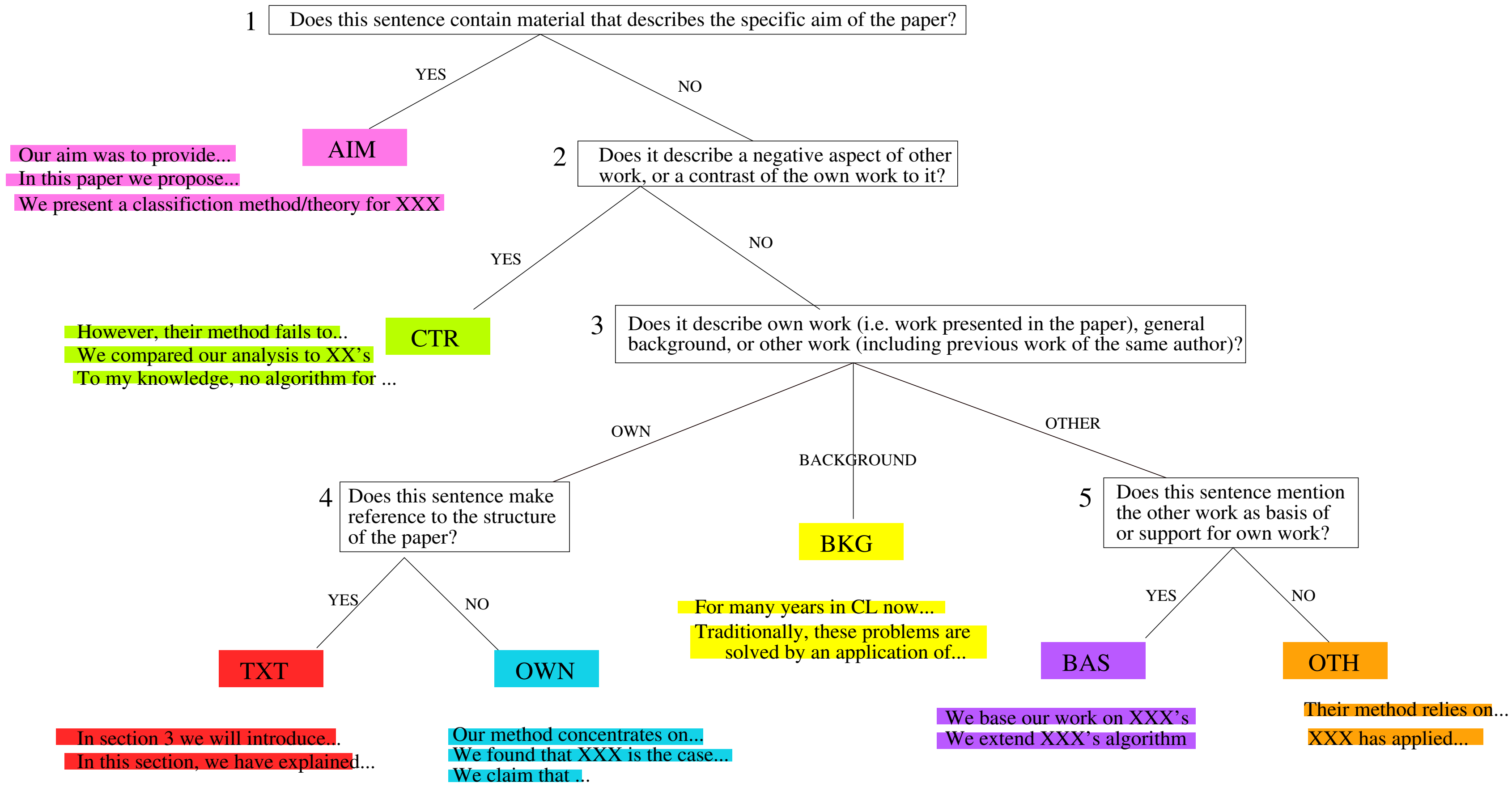
- sentence in the main body
- sentence in the abstract
- caption of a figure or a table
- figure, table, equation in running text
- example sentence (in linguistics papers)

as one and only one of the seven categories, using the decision tree on the other side to make your choice. Try not to leave anything uncoded. If you feel that more than one category applies to one entity, then choose the first one you come to in the decision tree. You should look at the surrounding context when making your choice. Try to annotate from the author's perspective, even if you do not agree with their portrayal of the situation.

When you are done with coding, please put a star next to the one single sentence in the main body of the text (not in the abstract!) that best expresses what the paper was about.

Some rules of thumb for assigning the categories:

- Not all papers have all categories.
- OWN, OTHER, BACKGROUND often come in chunks and there are many of them.
- CONTRAST, BASIS, AIM, TEXTUAL often come singly and they are rarer.


1 Does this sentence contain material that describes the specific aim of the paper?
YES
NO
AIM
Our aim was to provide...
In this paper we propose...
We present a classifiction method/theory for XXX
2 Does it describe a negative aspect of other work, or a contrast of the own work to it?
YES
NO
CTR
However, their method fails to...
We compared our analysis to XX's
To my knowledge, no algorithm for ...
3 Does it describe own work (i.e. work presented in the paper), general background, or other work (including previous work of the same author)?
OWN
BACKGROUND
OTHER
4 Does this sentence make reference to the structure of the paper?
BKG
For many years in CL now...
Traditionally, these problems are solved by an application of...
5 Does this sentence mention the other work as basis of or support for own work?
YES
NO
TXT
In section 3 we will introduce...
In this section, we have explained...
OWN
Our method concentrates on...
We found that XXX is the case...
We claim that ...
YES
NO
BAS
We base our work on XXX's
We extend XXX's algorithm
OTH
Their method relies on...
XXX has applied...

# Appendix D

# Lexical Resources

## D.1. Formulaic Patterns

| | |
|---|---|
| GENERAL_FORMULAIC | in @TRADITION_ADJ JJ ↑@WORK_NOUN |
| | in @TRADITION_ADJ used ↑@WORK_NOUN |
| | in @TRADITION_ADJ ↑@WORK_NOUN |
| | in @MANY JJ ↑@WORK_NOUN |
| | in @MANY ↑@WORK_NOUN |
| | in @BEFORE_ADJ JJ ↑@WORK_NOUN |
| | in @BEFORE_ADJ ↑@WORK_NOUN |
| | in other JJ ↑@WORK_NOUN |
| | in other ↑@WORK_NOUN |
| | in such ↑@WORK_NOUN |
| THEM_FORMULAIC | ↑according to CITE |
| | along the ↑lines of CITE |
| | ↑like CITE |
| | CITE ↑style |
| | a la ↑CITE |
| | CITE - ↑style |
| US_PREVIOUS_FORMULAIC | @SELF_NOM have ↑previously |
| | @SELF_NOM have ↑earlier |
| | @SELF_NOM have ↑elsewhere |
| | @SELF_NOM ↑elsewhere |
| | @SELF_NOM ↑previously |
| | @SELF_NOM ↑earlier |
| | ↑elsewhere @SELF_NOM |
| | ↑elswhere @SELF_NOM |
| | ↑elsewhere , @SELF_NOM |
| | ↑elswhere , @SELF_NOM |
| | presented ↑elswhere |
| | presented ↑elsewhere |
| | @SELF_NOM have shown ↑elsewhere |
| | @SELF_NOM have argued ↑elsewhere |
| | @SELF_NOM have shown ↑elswhere_NOM |
| | @SELF_NOM have argued ↑elswhere_NOM |
| | @SELF_NOM will show ↑elsewhere |
| | @SELF_NOM will show ↑elswhere |

| | |
|---|---|
| | @SELF_NOM will argue ↑elsewhere |
| | @SELF_NOM will argue ↑elswhere |
| | ↑elsewhere SELFCITE |
| | ↑elswhere SELFCITE |
| | in a @BEFORE_ADJ ↑@PRESENTATION_NOUN |
| | in an earlier ↑@PRESENTATION_NOUN |
| | another ↑@PRESENTATION_NOUN |
| TEXTSTRUCTURE_FORMULAIC | ↑then @SELF_NOM describe |
| | ↑then , @SELF_NOM describe |
| | ↑next @SELF_NOM describe |
| | ↑next , @SELF_NOM describe |
| | ↑finally @SELF_NOM describe |
| | ↑finally , @SELF_NOM describe |
| | ↑then @SELF_NOM present |
| | ↑then , @SELF_NOM present |
| | ↑next @SELF_NOM present |
| | ↑next , @SELF_NOM present |
| | ↑finally @SELF_NOM present |
| | ↑finally , @SELF_NOM present |
| | ↑briefly describe |
| | ↑briefly introduce |
| | ↑briefly present |
| | ↑briefly discuss |
| HERE_FORMULAIC | in this ↑@PRESENTATION_NOUN |
| | the present ↑@PRESENTATION_NOUN |
| | @SELF_NOM ↑here |
| | ↑here @SELF_NOM |
| | ↑here , @SELF_NOM |
| | @GIVEN ↑here |
| | @SELF_NOM ↑now |
| | ↑now @SELF_NOM |
| | ↑now , @SELF_NOM |
| | @GIVEN ↑now |
| | herein |
| METHOD_FORMULAIC | a new ↑@WORK_NOUN |
| | a novel ↑@WORK_NOUN |
| | a ↑@WORK_NOUN of |
| | an ↑@WORK_NOUN of |
| | a JJ ↑@WORK_NOUN of |
| | an JJ ↑@WORK_NOUN of |
| | a NN ↑@WORK_NOUN of |
| | an NN ↑@WORK_NOUN of |
| | a JJ NN ↑@WORK_NOUN of |
| | an JJ NN ↑@WORK_NOUN of |
| | a ↑@WORK_NOUN for |
| | an ↑@WORK_NOUN for |
| | a JJ ↑@WORK_NOUN for |
| | an JJ ↑@WORK_NOUN for |
| | a NN ↑@WORK_NOUN for |
| | an NN ↑@WORK_NOUN for |
| | a JJ NN ↑@WORK_NOUN for |
| | an JJ NN ↑@WORK_NOUN for |
| | ↑@WORK_NOUN designed to VV |

| | |
|---|---|
| | ↑@WORK_NOUN intended for |
| | ↑@WORK_NOUN for VV_ING |
| | ↑@WORK_NOUN for the NN |
| | ↑@WORK_NOUN designed to VV |
| | ↑@WORK_NOUN to the NN |
| | ↑@WORK_NOUN to NN |
| | ↑@WORK_NOUN to VV_ING |
| | ↑@WORK_NOUN for JJ VV_ING |
| | ↑@WORK_NOUN for the JJ NN |
| | ↑@WORK_NOUN to the JJ NN |
| | ↑@WORK_NOUN to JJ VV_ING |
| | the ↑problem of RB VV_ING |
| | the ↑problem of VV_ING |
| | the ↑problem of how to |
| CONTINUE_FORMULAIC | ↑following CITE |
| | ↑following the @WORK_NOUN of CITE |
| | ↑following the @WORK_NOUN given in CITE |
| | ↑following the @WORK_NOUN presented in CITE |
| | ↑following the @WORK_NOUN proposed in CITE |
| | ↑following the @WORK_NOUN discussed in CITE |
| | ↑adopt CITE ’s |
| | ↑starting point for @REFERENTIAL @WORK_NOUN |
| | ↑starting point for @SELF_POSS @WORK_NOUN |
| | as a ↑starting point |
| | as ↑starting point |
| | ↑use CITE ’s |
| | ↑base @SELF_POSS |
| | ↑supports @SELF_POSS |
| | ↑supports @OTHERS_POSS |
| | ↑support @OTHERS_POSS |
| | ↑support @SELF_POSS |
| | lends ↑support to @SELF_POSS |
| | lends ↑support to @OTHERS_POSS |
| CONTRAST_FORMULAIC | however, nevertheless, nonetheless, unfortunately, yet, although |
| GAP_FORMULAIC | as far as @SELF_NOM ↑know |
| | to @SELF_POSS ↑knowledge |
| | to the best of @SELF_POSS ↑knowledge |
| FUTURE_FORMULAIC | in the ↑future |
| | in the near ↑future |
| | ↑@FUTURE_ADJ @WORK_NOUN |
| | ↑@FUTURE_ADJ @AIM_NOUN |
| | ↑@FUTURE_ADJ development |
| | needs ↑further |
| | requires ↑further |
| | beyond the ↑scope |
| | ↑avenue for improvement |
| | ↑avenues for improvement |
| | ↑avenues for @FUTURE_ADJ improvement |
| | ↑areas for @FUTURE_ADJ improvement |
| | ↑areas for improvement |
| | ↑avenues of @FUTURE_ADJ research |
| | promising ↑avenue |
| | promising ↑avenues |

| | |
|---|---|
| SIMILARITY_FORMULAIC | along the same ↑lines<br>in a ↑similar vein<br>as in ↑@SELF_POSS<br>as in ↑CITE<br>as ↑did CITE<br>like in ↑CITE<br>↑like CITE 's<br>similarity with ↑CITE<br>similarity with ↑@SELF_POSS<br>similarity with ↑@OTHERS_POSS<br>↑similarity with @TRADITION_ADJ<br>↑similarity with @MANY<br>↑similarity with @BEFORE_ADJ<br>in analogy to ↑CITE<br>in analogy to ↑@SELF_POSS<br>in analogy to ↑@OTHERS_POSS<br>in ↑analogy to @TRADITION_ADJ<br>in ↑analogy to @MANY<br>in ↑analogy to @BEFORE_ADJ<br>↑similar to that described here<br>↑similar to that of<br>↑similar to those of<br>↑similar to CITE<br>↑similar to @SELF_ACC<br>↑similar to @SELF_POSS<br>↑similar to @OTHERS_ACC<br>↑similar to @TRADITION_ADJ<br>↑similar to @MANY<br>↑similar to @BEFORE_ADJ<br>↑similar to @OTHERS_POSS<br>↑similar to CITE<br>a ↑similar NN to @SELF_POSS<br>a ↑similar NN to @OTHERS_POSS<br>a ↑similar NN to CITE<br>↑analogous to that described here<br>↑analogous to CITE<br>↑analogous to @SELF_ACC<br>↑analogous to @SELF_POSS<br>↑analogous to @OTHERS_ACC<br>↑analogous to @TRADITION_ADJ<br>↑analogous to @MANY<br>↑analogous to @BEFORE_ADJ<br>↑analogous to @OTHERS_POSS<br>↑analogous to CITE<br>the ↑same NN as @SELF_POSS<br>the ↑same NN as @OTHERS_POSS<br>the ↑same NN as CITE<br>the ↑same as @SELF_POSS<br>the ↑same as @OTHERS_POSS<br>the ↑same as CITE<br>in ↑common with @OTHERS_POSS<br>in ↑common with @SELF_POSS<br>in ↑common with @TRADITION_ADJ |

| | |
|---|---|
| | in ↑common with @MANY |
| | in ↑common with @BEFORE_ADJ |
| | most ↑relevant to @SELF_POSS |
| COMPARISON_FORMULAIC | ↑against CITE |
| | ↑against @SELF_ACC |
| | ↑against @SELF_POSS |
| | ↑against @OTHERS_ACC |
| | ↑against @OTHERS_POSS |
| | ↑against @BEFORE_ADJ @WORK_NOUN |
| | ↑against @MANY @WORK_NOUN |
| | ↑against @TRADITION_ADJ @WORK_NOUN |
| | ↑than CITE |
| | ↑than @SELF_ACC |
| | ↑than @SELF_POSS |
| | ↑than @OTHERS_ACC |
| | ↑than @OTHERS_POSS |
| | ↑than @TRADITION_ADJ @WORK_NOUN |
| | ↑than @BEFORE_ADJ @WORK_NOUN |
| | ↑than @MANY @WORK_NOUN |
| | point of ↑departure from @SELF_POSS |
| | points of ↑departure from @OTHERS_POSS |
| | ↑advantage over @OTHERS_ACC |
| | ↑advantage over @TRADITION_ADJ |
| | ↑advantage over @MANY @WORK_NOUN |
| | ↑advantage over @BEFORE_ADJ @WORK_NOUN |
| | ↑advantage over @OTHERS_POSS |
| | ↑advantage over CITE |
| | ↑advantage to @OTHERS_ACC |
| | ↑advantage to @OTHERS_POSS |
| | ↑advantage to CITE |
| | ↑advantage to @TRADITION_ADJ |
| | ↑advantage to @MANY @WORK_NOUN |
| | ↑advantage to @BEFORE_ADJ @WORK_NOUN |
| | ↑advantages over @OTHERS_ACC |
| | ↑advantages over @TRADITION_ADJ |
| | ↑advantages over @MANY @WORK_NOUN |
| | ↑advantages over @BEFORE_ADJ @WORK_NOUN |
| | ↑advantages over @OTHERS_POSS |
| | ↑advantages over CITE |
| | ↑advantages to @OTHERS_ACC |
| | ↑advantages to @OTHERS_POSS |
| | ↑advantages to CITE |
| | ↑advantages to @TRADITION_ADJ |
| | ↑advantages to @MANY @WORK_NOUN |
| | ↑advantages to @BEFORE_ADJ @WORK_NOUN |
| | ↑benefit over @OTHERS_ACC |
| | ↑benefit over @OTHERS_POSS |
| | ↑benefit over CITE |
| | ↑benefit over @TRADITION_ADJ |
| | ↑benefit over @MANY @WORK_NOUN |
| | ↑benefit over @BEFORE_ADJ @WORK_NOUN |
| | ↑difference to CITE |
| | ↑difference to @TRADITION_ADJ |

↑difference to CITE
↑difference to @TRADITION_ADJ
↑difference to @MANY @WORK_NOUN
↑difference to @BEFORE_ADJ @WORK_NOUN
↑difference to @OTHERS_ACC
↑difference to @OTHERS_POSS
↑difference to @SELF_ACC
↑difference to @SELF_POSS
↑differences to CITE
↑differences to @TRADITION_ADJ
↑differences to @MANY @WORK_NOUN
↑differences to @BEFORE_ADJ @WORK_NOUN
↑differences to @OTHERS_ACC
↑differences to @OTHERS_POSS
↑differences to @SELF_ACC
↑differences to @SELF_POSS
↑difference between CITE
↑difference between @TRADITION_ADJ
↑difference between @MANY @WORK_NOUN
↑difference between @BEFORE_ADJ @WORK_NOUN
↑difference between @OTHERS_ACC
↑difference between @OTHERS_POSS
↑difference between @SELF_ACC
↑difference between @SELF_POSS
↑differences between CITE
↑differences between @TRADITION_ADJ
↑differences between @MANY @WORK_NOUN
↑differences between @BEFORE_ADJ @WORK_NOUN
↑differences between @OTHERS_ACC
↑differences between @OTHERS_POSS
↑differences between @SELF_ACC
↑differences between @SELF_POSS
↑contrast with CITE
↑contrast with @TRADITION_ADJ
↑contrast with @MANY @WORK_NOUN
↑contrast with @BEFORE_ADJ @WORK_NOUN
↑contrast with @OTHERS_ACC
↑contrast with @OTHERS_POSS
↑contrast with @SELF_ACC
↑contrast with @SELF_POSS
↑unlike @SELF_ACC
↑unlike @SELF_POSS
↑unlike CITE
↑unlike @TRADITION_ADJ
↑unlike @BEFORE_ADJ @WORK_NOUN
↑unlike @MANY @WORK_NOUN
↑unlike @OTHERS_ACC
↑unlike @OTHERS_POSS
in ↑contrast to @SELF_ACC
in ↑contrast to @SELF_POSS
in ↑contrast to CITE
in ↑contrast to @TRADITION_ADJ
in ↑contrast to @MANY @WORK_NOUN

| | |
|---|---|
| | in ↑contrast to @BEFORE_ADJ @WORK_NOUN |
| | in ↑contrast to @OTHERS_ACC |
| | in ↑contrast to @OTHERS_POSS |
| | as ↑opposed to @SELF_ACC |
| | as ↑opposed to @SELF_POSS |
| | as ↑opposed to CITE |
| | as ↑opposed to @TRADITION_ADJ |
| | as ↑opposed to @MANY @WORK_NOUN |
| | as ↑opposed to @BEFORE_ADJ @WORK_NOUN |
| | as ↑opposed to @OTHERS_ACC |
| | as ↑opposed to @OTHERS_POSS |
| | ↑contrary to @SELF_ACC |
| | ↑contrary to @SELF_POSS |
| | ↑contrary to CITE |
| | ↑contrary to @TRADITION_ADJ |
| | ↑contrary to @MANY @WORK_NOUN |
| | ↑contrary to @BEFORE_ADJ @WORK_NOUN |
| | ↑contrary to @OTHERS_ACC |
| | ↑contrary to @OTHERS_POSS |
| | ↑whereas @SELF_ACC |
| | ↑whereas @SELF_POSS |
| | ↑whereas CITE |
| | ↑whereas @TRADITION_ADJ |
| | ↑whereas @BEFORE_ADJ @WORK_NOUN |
| | ↑whereas @MANY @WORK_NOUN |
| | ↑whereas @OTHERS_ACC |
| | ↑whereas @OTHERS_POSS |
| | ↑compared to @SELF_ACC |
| | ↑compared to @SELF_POSS |
| | ↑compared to CITE |
| | ↑compared to @TRADITION_ADJ |
| | ↑compared to @BEFORE_ADJ @WORK_NOUN |
| | ↑compared to @MANY @WORK_NOUN |
| | ↑compared to @OTHERS_ACC |
| | ↑compared to @OTHERS_POSS |
| | in ↑comparison to @SELF_ACC |
| | in ↑comparison to @SELF_POSS |
| | in ↑comparison to CITE |
| | in ↑comparison to @TRADITION_ADJ |
| | in ↑comparison to @MANY @WORK_NOUN |
| | in ↑comparison to @BEFORE_ADJ @WORK_NOUN |
| | in ↑comparison to @OTHERS_ACC |
| | in ↑comparison to @OTHERS_POSS |
| | ↑while @SELF_NOM |
| | ↑while @SELF_POSS |
| | ↑while CITE |
| | ↑while @TRADITION_ADJ |
| | ↑while @BEFORE_ADJ @WORK_NOUN |
| | ↑while @MANY @WORK_NOUN |
| | ↑while @OTHERS_NOM |
| | ↑while @OTHERS_POSS |
| AFFECT_FORMULAIC | hopefully |
| | thankfully |

| | |
|---|---|
| | fortunately |
| | unfortunately |
| GOOD_FORMULAIC | @POS_ADJ |
| BAD_FORMULAIC | @NEG_ADJ |
| TRADITION_FORMULAIC | @TRADITIONAL_ADJ |
| IN_ORDER_TO_FORMULAIC | in ↑order to |
| DETAIL_FORMULAIC | @SELF_NOM have ↑also |
| | @SELF_NOM ↑also |
| | this @PRESENTATION_NOUN ↑also |
| | this @PRESENTATION_NOUN has ↑also |
| NO_TEXTSTRUCTURE_FORMULAIC | ( ↑TXT_NOUN CREF ) |
| | as explained in ↑@TXT_NOUN CREF |
| | as explained in the @BEFORE_ADJ ↑@TXT_NOUN |
| | as ↑@GIVEN earlier in this @TXT_NOUN |
| | as ↑@GIVEN below |
| | as @GIVEN in ↑@TXT_NOUN CREF |
| | as @GIVEN in the @BEFORE_ADJ ↑@TXT_NOUN |
| | as @GIVEN in the next ↑@TXT_NOUN |
| | NN @GIVEN in ↑@TXT_NOUN CREF |
| | NN @GIVEN in the @BEFORE_ADJ ↑@TXT_NOUN |
| | NN @GIVEN in the next ↑@TXT_NOUN |
| | NN @GIVEN ↑below |
| | cf. ↑@TXT_NOUN CREF |
| | cf. ↑@TXT_NOUN below |
| | cf. the ↑@TXT_NOUN below |
| | cf. the @BEFORE_ADJ ↑@TXT_NOUN |
| | cf. ↑@TXT_NOUN above |
| | cf. the ↑@TXT_NOUN above |
| | e. g. , ↑@TXT_NOUN CREF |
| | e. g , ↑@TXT_NOUN CREF |
| | e. g. ↑@TXT_NOUN CREF |
| | e. g ↑@TXT_NOUN CREF |
| | compare ↑@TXT_NOUN CREF |
| | compare ↑@TXT_NOUN below |
| | compare the ↑@TXT_NOUN below |
| | compare the @BEFORE_ADJ ↑@TXT_NOUN |
| | compare ↑@TXT_NOUN above |
| | compare the ↑@TXT_NOUN above |
| | see ↑@TXT_NOUN CREF |
| | see the @BEFORE_ADJ ↑@TXT_NOUN |
| | recall from the @BEFORE_ADJ ↑@TXT_NOUN |
| | recall from the ↑@TXT_NOUN above |
| | recall from ↑@TXT_NOUN CREF |
| | @SELF_NOM shall see ↑below |
| | @SELF_NOM will see ↑below |
| | @SELF_NOM shall see in the ↑next @TXT_NOUN |
| | @SELF_NOM will see in the ↑next @TXT_NOUN |
| | @SELF_NOM shall see in ↑@TXT_NOUN CREF |
| | @SELF_NOM will see in ↑@TXT_NOUN CREF |
| | example in ↑@TXT_NOUN CREF |
| | example CREF in ↑@TXT_NOUN CREF |
| | examples CREF and CREF in ↑@TXT_NOUN CREF |
| | examples in ↑@TXT_NOUN CREF |

# D.2. Agent Patterns

| | |
|---|---|
| US_AGENT | @SELF_NOM |
| | @SELF_POSS JJ ↑@WORK_NOUN |
| | @SELF_POSS JJ ↑@PRESENTATION_NOUN |
| | @SELF_POSS JJ ↑@ARGUMENTATION_NOUN |
| | @SELF_POSS JJ ↑@SOLUTION_NOUN |
| | @SELF_POSS JJ ↑@RESULT_NOUN |
| | @SELF_POSS ↑@WORK_NOUN |
| | @SELF_POSS ↑@PRESENTATION_NOUN |
| | @SELF_POSS ↑@ARGUMENTATION_NOUN |
| | @SELF_POSS ↑@SOLUTION_NOUN |
| | @SELF_POSS ↑@RESULT_NOUN |
| | ↑@WORK_NOUN @GIVEN here |
| | ↑@WORK_NOUN @GIVEN below |
| | ↑@WORK_NOUN @GIVEN in this @PRESENTATION_NOUN |
| | ↑@WORK_NOUN @GIVEN in @SELF_POSS @PRESENTATION_NOUN |
| | the ↑@SOLUTION_NOUN @GIVEN here |
| | the ↑@SOLUTION_NOUN @GIVEN in this @PRESENTATION_NOUN |
| | the first ↑author |
| | the second ↑author |
| | the third ↑author |
| | one of the ↑authors |
| | one of ↑us |
| REF_US_AGENT | this ↑@PRESENTATION_NOUN |
| | the present ↑@PRESENTATION_NOUN |
| | the current ↑@PRESENTATION_NOUN |
| | the present JJ ↑@PRESENTATION_NOUN |
| | the current JJ ↑@PRESENTATION_NOUN |
| | the ↑@WORK_NOUN @GIVEN |
| OUR_AIM_AGENT | @SELF_POSS ↑@AIM_NOUN |
| | the point of this ↑@PRESENTATION_NOUN |
| | the ↑@AIM_NOUN of this @PRESENTATION_NOUN |
| | the ↑@AIM_NOUN of the @GIVEN @WORK_NOUN |
| | the ↑@AIM_NOUN of @SELF_POSS @WORK_NOUN |
| | the ↑@AIM_NOUN of @SELF_POSS @PRESENTATION_NOUN |
| | the most important feature of ↑@SELF_POSS @WORK_NOUN |
| | contribution of this ↑@PRESENTATION_NOUN |
| | contribution of the @GIVEN ↑@WORK_NOUN |
| | contribution of ↑@SELF_POSS @WORK_NOUN |
| | the question @GIVEN in this ↑PRESENTATION_NOUN |
| | the question @GIVEN ↑here |
| | @SELF_POSS @MAIN ↑@AIM_NOUN |
| | @SELF_POSS ↑@AIM_NOUN in this @PRESENTATION_NOUN |
| | @SELF_POSS ↑@AIM_NOUN here |
| | the JJ point of this ↑@PRESENTATION_NOUN |
| | the JJ purpose of this ↑@PRESENTATION_NOUN |
| | the JJ ↑@AIM_NOUN of this @PRESENTATION_NOUN |
| | the JJ ↑@AIM_NOUN of the @GIVEN @WORK_NOUN |
| | the JJ ↑@AIM_NOUN of @SELF_POSS @WORK_NOUN |
| | the JJ ↑@AIM_NOUN of @SELF_POSS @PRESENTATION_NOUN |
| | the JJ question @GIVEN in this ↑PRESENTATION_NOUN |

| | |
|---|---|
| | the JJ question @GIVEN ↑here |
| AIM_REF_AGENT | its ↑@AIM_NOUN |
| | its JJ ↑@AIM_NOUN |
| | @REFERENTIAL JJ ↑@AIM_NOUN |
| | contribution of this ↑@WORK_NOUN |
| | the most important feature of this ↑@WORK_NOUN |
| | feature of this ↑@WORK_NOUN |
| | the ↑@AIM_NOUN |
| | the JJ ↑@AIM_NOUN |
| US_PREVIOUS_AGENT | SELFCITE |
| | this @BEFORE_ADJ ↑@PRESENTATION_NOUN |
| | @SELF_POSS @BEFORE_ADJ ↑@PRESENTATION_NOUN |
| | @SELF_POSS @BEFORE_ADJ ↑@WORK_NOUN |
| | in ↑SELFCITE , @SELF_NOM |
| | in ↑SELFCITE @SELF_NOM |
| | the ↑@WORK_NOUN @GIVEN in SELFCITE |
| REF_AGENT | @REFERENTIAL JJ ↑@WORK_NOUN |
| | @REFERENTIAL ↑@WORK_NOUN |
| | this sort of ↑@WORK_NOUN |
| | this kind of ↑@WORK_NOUN |
| | this type of ↑@WORK_NOUN |
| | the current JJ ↑@WORK_NOUN |
| | the current ↑@WORK_NOUN |
| | the ↑@WORK_NOUN |
| | the ↑@PRESENTATION_NOUN |
| | the ↑author |
| | the ↑authors |
| THEM_PRONOUN_AGENT | @OTHERS_NOM |
| THEM_AGENT | CITE |
| | CITE 's NN |
| | CITE 's ↑@PRESENTATION_NOUN |
| | CITE 's ↑@WORK_NOUN |
| | CITE 's ↑@ARGUMENTATION_NOUN |
| | CITE 's JJ ↑@PRESENTATION_NOUN |
| | CITE 's JJ ↑@WORK_NOUN |
| | CITE 's JJ ↑@ARGUMENTATION_NOUN |
| | the CITE ↑@WORK_NOUN |
| | the ↑@WORK_NOUN @GIVEN in CITE |
| | the ↑@WORK_NOUN of CITE |
| | @OTHERS_POSS ↑@PRESENTATION_NOUN |
| | @OTHERS_POSS ↑@WORK_NOUN |
| | @OTHERS_POSS ↑@RESULT_NOUN |
| | @OTHERS_POSS ↑@ARGUMENTATION_NOUN |
| | @OTHERS_POSS ↑@SOLUTION_NOUN |
| | @OTHERS_POSS JJ ↑@PRESENTATION_NOUN |
| | @OTHERS_POSS JJ ↑@WORK_NOUN |
| | @OTHERS_POSS JJ ↑@RESULT_NOUN |
| | @OTHERS_POSS JJ ↑@ARGUMENTATION_NOUN |
| | @OTHERS_POSS JJ ↑@SOLUTION_NOUN |
| GAP_AGENT | none of these ↑@WORK_NOUN |
| | none of those ↑@WORK_NOUN |
| | no ↑@WORK_NOUN |
| | no JJ ↑@WORK_NOUN |

| | |
|---|---|
| | none of these ↑@PRESENTATION_NOUN |
| | none of those ↑@PRESENTATION_NOUN |
| | no ↑@PRESENTATION_NOUN |
| | no JJ ↑@PRESENTATION_NOUN |
| GENERAL_AGENT | @TRADITION_ADJ JJ ↑@WORK_NOUN |
| | @TRADITION_ADJ used ↑@WORK_NOUN |
| | @TRADITION_ADJ ↑@WORK_NOUN |
| | @MANY JJ ↑@WORK_NOUN |
| | @MANY ↑@WORK_NOUN |
| | @BEFORE_ADJ JJ ↑@WORK_NOUN |
| | @BEFORE_ADJ ↑@WORK_NOUN |
| | @BEFORE_ADJ JJ ↑@PRESENTATION_NOUN |
| | @BEFORE_ADJ ↑@PRESENTATION_NOUN |
| | other JJ ↑@WORK_NOUN |
| | other ↑@WORK_NOUN |
| | such ↑@WORK_NOUN |
| | these JJ ↑@PRESENTATION_NOUN |
| | these ↑@PRESENTATION_NOUN |
| | those JJ ↑@PRESENTATION_NOUN |
| | those ↑@PRESENTATION_NOUN |
| | @REFERENTIAL ↑authors |
| | @MANY ↑authors |
| | ↑researchers in @DISCIPLINE |
| | @PROFESSIONAL_NOUN |
| PROBLEM_AGENT | @REFERENTIAL JJ ↑@PROBLEM_NOUN |
| | @REFERENTIAL ↑@PROBLEM_NOUN |
| | the ↑@PROBLEM_NOUN |
| SOLUTION_AGENT | @REFERENTIAL JJ ↑@SOLUTION_NOUN |
| | @REFERENTIAL ↑@SOLUTION_NOUN |
| | the ↑@SOLUTION_NOUN |
| | the JJ ↑@SOLUTION_NOUN |
| TEXTSTRUCTURE_AGENT | ↑@TXT_NOUN CREF |
| | ↑@TXT_NOUN CREF and CREF |
| | this ↑@TXT_NOUN |
| | next ↑@TXT_NOUN |
| | next CD ↑@TXT_NOUN |
| | concluding ↑@TXT_NOUN |
| | @BEFORE_ADJ ↑@TXT_NOUN |
| | ↑@TXT_NOUN above |
| | ↑@TXT_NOUN below |
| | following ↑@TXT_NOUN |
| | remaining ↑@TXT_NOUN |
| | subsequent ↑@TXT_NOUN |
| | following CD ↑@TXT_NOUN |
| | remaining CD ↑@TXT_NOUN |
| | subsequent CD ↑@TXT_NOUN |
| | ↑@TXT_NOUN that follow |
| | rest of this ↑@PRESENTATION_NOUN |
| | remainder of this ↑@PRESENTATION_NOUN |
| | in ↑@TXT_NOUN CREF , @SELF_NOM |
| | in this ↑@TXT_NOUN , @SELF_NOM |
| | in the next ↑@TXT_NOUN , @SELF_NOM |
| | in @BEFORE_ADJ ↑@TXT_NOUN , @SELF_NOM |

in the @BEFORE_ADJ ↑@TXT_NOUN , @SELF_NOM
in the ↑@TXT_NOUN above , @SELF_NOM
in the ↑@TXT_NOUN below , @SELF_NOM
in the following ↑@TXT_NOUN , @SELF_NOM
in the remaining ↑@TXT_NOUN , @SELF_NOM
in the subsequent ↑@TXT_NOUN , @SELF_NOM
in the ↑@TXT_NOUN that follow , @SELF_NOM
in the rest of this ↑@PRESENTATION_NOUN , @SELF_NOM
in the remainder of this ↑@PRESENTATION_NOUN , @SELF_NOM
↑below , @SELF_NOM
the ↑@AIM_NOUN of this @TXT_NOUN

# D.3. Action Lexicon

| | |
|---|---|
| AFFECT | afford, believe, decide, feel, hope, imagine, regard, trust, think |
| ARGUMENTATION | agree, accept, advocate, argue, claim, conclude, comment, defend, embrace, hypothesize, imply, insist, posit, postulate, reason, recommend, speculate, stipulate, suspect |
| AWARE | be unaware, be familiar with, be aware, be not aware, know of |
| BETTER_SOLUTION | boost, enhance, defeat, improve, go beyond, perform better, outperform, outweigh, surpass |
| CHANGE | adapt, adjust, augment, combine, change, decrease, elaborate, expand, extend, derive, incorporate, increase, manipulate, modify, optimize, optimise, refine, render, replace, revise, substitute, tailor, upgrade |
| COMPARISON | compare, compete, evaluate, test |
| CONTINUE | adopt, agree with CITE, base, be based on, be derived from, be originated in, be inspired by, borrow, build on, follow CITE, originate from, originate in, side with |
| CONTRAST | be different from, be distinct from, conflict, contrast, clash, differ from, distinguish @RFX, differentiate, disagree, disagreeing, dissent, oppose |
| FUTURE_INTEREST | plan on, plan to, expect to, intend to |
| INTEREST | aim, ask @SELF_RFX, ask @OTHERS_RFX, address, attempt, be concerned, be interested, be motivated, concern, concern @SELF_ACC, concern @OTHERS_ACC, consider, concentrate on, explore, focus, intend to, like to, look at how, motivate @SELF_ACC, motivate @OTHERS_ACC, pursue, seek, study, try, target, want, wish, wonder |
| NEED | be dependent on, be reliant on, depend on, lack, need, necessitate, require, rely on |
| PRESENTATION | describe, discuss, give, introduce, note, notice, point out, present, propose, put forward, recapitulate, remark, report, say, show, sketch, state, suggest, talk about |
| PROBLEM | abound, aggravate, arise, be cursed, be incapable of, be forced to, be limited to, be problematic, be restricted to, be troubled, be unable to, contradict, damage, degrade, degenerate, fail, fall prey, fall short, force @SELF_ACC, force @OTHERS_ACC, hinder, impair, impede, inhibit, misclassify, misjudge, mistake, misuse, neglect, obscure, overestimate, over-estimate, overfit, over-fit, overgeneralize, over-generalize, overgeneralise, over-generalise, overgenerate, over-generate, overlook, pose, plague, preclude, prevent, remain, resort to, restrain, run into, settle for, spoil, suffer from, threaten, thwart, underestimate, under-estimate, undergenerate, under-generate, violate, waste, worsen |
| RESEARCH | apply, analyze, analyse, build, calculate, categorize, categorise, characterize, characterise, choose, check, classify, collect, compose, compute, conduct, confirm, construct, count, define, delineate, detect, determine, equate, estimate, examine, expect, formalize, formalise, formulate, gather, identify, implement, |

indicate, inspect, integrate, interpret, investigate, isolate, maximize, maximise, measure, minimize, minimise, observe, predict, realize, realise, reconfirm, simulate, select, specify, test, verify

| | |
|---|---|
| SIMILAR | bear comparison, be analogous to, be alike, be related to, be closely related to, be reminiscent of, be the same as, be similar to, be in a similar vein to, have much in common with, have a lot in common with, pattern with, resemble |
| SOLUTION | accomplish, account for, achieve, apply to, answer, alleviate, allow for, allow @SELF_ACC, allow @OTHERS_ACC, avoid, benefit, capture, clarify, circumvent, contribute, cope with, cover, cure, deal with, demonstrate, develop, devise, discover, elucidate, escape, explain, fix, gain, go a long way, guarantee, handle, help, implement, justify, lend itself, make progress, manage, mend, mitigate, model, obtain, offer, overcome, perform, preserve, prove, provide, realize, realise, rectify, refrain from, remedy, resolve, reveal, scale up, sidestep, solve, succeed, tackle, take care of, take into account, treat, warrant, work well, yield |
| TEXTSTRUCTURE | begin by, illustrate, conclude by, organize, organise, outline, return to, review, start by, structure, summarize, summarise, turn to |
| USE | apply, employ, use, make use, utilize |

# D.4. Concept Lexicon

| | |
|---|---|
| NEGATION | no, not, nor, non, neither, none, never, aren't, can't, cannot, hadn't, hasn't, haven't, isn't, didn't, don't, doesn't, n't, wasn't, weren't, nothing, nobody, less, least, little, scant, scarcely, rarely, hardly, few, rare, unlikely |
| 3RD PERSON PRONOUN (NOM) | they, he, she, theirs, hers, his |
| 3RD PERSON PRONOUN (ACC) | her, him, them |
| 3RD POSS PRONOUN | their, his, her |
| 3RD PERSON REFLEXIVE | themselves, himself, herself |
| 1ST PERSON PRONOUN (NOM) | we, i, ours, mine |
| 1ST PERSON PRONOUN (ACC) | us, me |
| 1ST POSS PRONOUN | my, our |
| 1ST PERSON REFLEXIVE | ourselves, myself |
| REFERENTIAL | this, that, those, these |
| REFLEXIVE | itself ourselves, myself, themselves, himself, herself |
| QUESTION | ?, how, why, whether, wonder |
| GIVEN | noted, mentioned, addressed, illustrated, described, discussed, given, outlined, presented, proposed, reported, shown, taken |
| PROFESSIONALS | collegues, community, computer scientists, computational linguists, discourse analysts, expert, investigators, linguists, logicians, philosophers, psycholinguists, psychologists, researchers, scholars, semanticists, scientists |
| DISCIPLINE | computer science, computer linguistics, computational linguistics, discourse analysis, logics, linguistics, psychology, psycholinguistics, philosophy, semantics, several disciplines, various disciplines |
| TEXT_NOUN | paragraph, section, subsection, chapter |
| SIMILAR_NOUN | analogy, similarity |
| COMPARISON_NOUN | accuracy, baseline, comparison, competition, evaluation, inferiority, measure, measurement, performance, precision, optimum, recall, superiority |
| CONTRAST_NOUN | contrast, conflict, clash, clashes, difference, point of departure |
| AIM_NOUN | aim, goal, intention, objective, purpose, task, theme, topic |
| ARGUMENTATION_NOUN | assumption, belief, hypothesis, hypotheses, claim, conclusion, confirmation, opinion, recommendation, stipulation, view |
| PROBLEM_NOUN | Achilles heel, caveat, challenge, complication, contradiction, damage, danger, deadlock, defect, detriment, difficulty, dilemma, disadvantage, disregard, doubt, downside, drawback, error, failure, fault, foil, flaw, handicap, hindrance, hurdle, ill, inflexibility, impediment, imperfection, intractability, inefficiency, inadequacy, inability, lapse, limitation, malheur, mishap, mischance, mistake, obstacle, oversight, pitfall, problem, shortcoming, threat, trouble, vulnerability, absence, dearth, deprivation, lack, loss, fraught, proliferation, spate |
| QUESTION_NOUN | question, conundrum, enigma, paradox, phenomena, phenomenon, puzzle, riddle |
| SOLUTION_NOUN | answer, accomplishment, achievement, advantage, benefit, breakthrough, contribution, explanation, idea, improvement, innovation, insight, justification, proposal, proof, remedy, solution, success, triumph, verification, victory |

| | |
|---|---|
| INTEREST_NOUN | attention, quest |
| RESEARCH_NOUN | evidence, experiment, finding, progress, observation, outcome, result |
| CHANGE_NOUN | alternative, adaptation, extension, development, modification, refinement, version, variant, variation |
| PRESENTATION_NOUN | article, draft, paper, project, report, study |
| NEED_NOUN | necessity, motivation |
| WORK_NOUN | account, algorithm, analysis, analyses, approach, approaches, application, architecture, characterization, characterisation, component, design, extension, formalism, formalization, formalisation, framework, implementation, investigation, machinery, method, methodology, model, module, moduls, process, procedure, program, prototype, research, researches, strategy, system, technique, theory, tool, treatment, work |
| TRADITION_NOUN | acceptance, community, convention, disciples, disciplines, folklore, literature, mainstream, school, tradition, textbook |
| CHANGE_ADJ | alternate, alternative |
| GOOD_ADJ | adequate, advantageous, appealing, appropriate, attractive, automatic, beneficial, capable, cheerful, clean, clear, compact, compelling, competitive, comprehensive, consistent, convenient, convincing, constructive, correct, desirable, distinctive, efficient, elegant, encouraging, exact, faultless, favourable, feasible, flawless, good, helpful, impeccable, innovative, insightful, intensive, meaningful, neat, perfect, plausible, positive, polynomial, powerful, practical, preferable, precise, principled, promising, pure, realistic, reasonable, reliable, right, robust, satisfactory, simple, sound, successful, sufficient, systematic, tractable, usable, useful, valid, unlimited, well worked out, well, enough |
| BAD_ADJ | absent, ad-hoc, adhoc, ad hoc, annoying, ambiguous, arbitrary, awkward, bad, brittle, brute-force, brute force, careless, confounding, contradictory, defect, defunct, disturbing, elusive, erraneous, expensive, exponential, false, fallacious, frustrating, haphazard, ill-defined, imperfect, impossible, impractical, imprecise, inaccurate, inadequate, inappropriate, incomplete, incomprehensible, inconclusive, incorrect, inelegant, inefficient, inexact, infeasible, infelicitous, inflexible, implausible, inpracticable, improper, insufficient, intractable, invalid, irrelevant, labour-intensive, labor-intensive, labour intensive, labor intensive, limited-coverage, limited coverage, limited, limiting, meaningless, modest, misguided, misleading, nonexistent, NP-hard, NP-complete, NP hard, NP complete, questionable, pathological, poor, prone, protracted, restricted, scarce, simplistic, suspect, time-consuming, time consuming, toy, unacceptable, unaccounted for, unaccounted-for, unaccounted, unattractive, unavailable, unavoidable, unclear, uncomfortable, unexplained, undecidable, undesirable, unfortunate, uninnovative, uninterpretable, unjustified, unmotivated, unnatural, unnecessary, unorthodox, unpleasant, unpractical, unprincipled, unreliable, unsatisfactory, unsound, unsuccessful, unsuited, unsystematic, untractable, unwanted, unwelcome, useless, vulnerable, weak, wrong, too, overly, only |
| BEFORE_ADJ | earlier, past, previous, prior |
| CONTRAST_ADJ | different, distinguishing, contrary, competing, rival |
| TRADITION_ADJ | better known, better-known, cited, classic, common, conventional, current, customary, established, existing, extant, available, favourite, fashionable, general, obvious, long-standing, mainstream, modern, naive, orthodox, popular, prevailing, prevalent, published, quoted, seminal, standard, textbook, traditional, trivial, typical, well-established, well-known, widely-assumed, unanimous, usual |

| | |
|---|---|
| MANY | a number of, a body of, a substantial number of, a substantial body of, most, many, several, various |
| COMPARISON_ADJ | evaluative, superior, inferior, optimal, better, best, worse, worst, greater, larger, faster, weaker, stronger |
| PROBLEM_ADJ | demanding, difficult, hard, non-trivial, nontrivial |
| RESEARCH_ADJ | empirical, experimental, exploratory, ongoing, quantitative, qualitative, preliminary, statistical, underway |
| AWARE_ADJ | unnoticed, understood, unexplored |
| NEED_ADJ | necessary |
| NEW_ADJ | new, novel,state-of-the-art, state of the art, leading-edge, leading edge, enhanced |
| FUTURE_ADJ | further, future |
| MAIN_ADJ | main, key, basic, central, crucial, essential, eventual, fundamental, great, important, key, largest, main, major, overall, primary, principle, serious, substantial, ultimate |

# Index of Citations