Human annotation of lexical chains: coverage and agreement measures

Bill Hollingsworth
University of Cambridge Computer Laboratory
william.hollingsworth@cl.cam.ac.uk

Simone Teufel
University of Cambridge Computer Laboratory
simone.teufel@cl.cam.ac.uk

ABSTRACT

Lexical chains have been successfully used in several previous applications, e.g. topic segmentation and summarization. In this paper, we address the problem of how to directly evaluate the quality of lexical chains, in comparison to a human gold standard. This is in contrast to previous work, where the formal evaluation either relied on a word sense disambiguation task or concentrated on the final application result (the summary or the text segmentation), rather than the lexical chains themselves. We present a small user study of human annotation of lexical chains, and a set of measures to measure how much agreement between sets of lexical chains there is. We also perform a small meta-evaluation to compare the best of these metrics, a partial overlap measure, to rankings of chains derived by introspection, which shows that our measure agrees reasonably well with intuition. We also describe our algorithm for chain creation, which varies from previous work in several aspects (for instance the fact that it allows for adjective attribution), and report its agreement with our human annotators in terms of our new measure.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCENatural Language Processing

General Terms

Lexical Chains, Human Annotation

Keywords

Adjectives, Lexical Cohesion, Similarity Measures

1. INTRODUCTION

An algorithm for creating lexical chains was first proposed by Morris and Hirst [11] and relies on the theory of lexical cohesion [4]. A lexical chain is a collection of terms that are related within a text by lexico-semantic relations, such as synonymy or similar relations.

Lexical chains have been used in various applications, such as automatic text summarization [1, 13], text segmentation [5, 14], correction of malapropisms [6], and automatic generation of hypertext links [3]. The exact mechanisms by which lexical chains are used in these approaches differs from application to application. In Barzilay et al, for instance, particularly relevant sentences useful for presentation in a summary are defined as those that contain many intersecting chains. In [5], a gap between pseudo-sentences is more likely to be a topic shift if the number of lexical chains spanning the gap is low. Importantly, the user never sees the lexical chains directly; end evaluation of these approaches is performed in terms of either a word sense disambiguation task [1, 6] or the end product (summaries, text segmentation, hypertext output). The quality of the lexical chains themselves is thus typically not formally evaluated in the existing work.

We propose lexical chains as a framework for a different task, automatic text skimming. The goal of an automatic skimmer is to provide an online user with the ability to browse the text in a document in a way similar to that of a reader of a printed document. This would allow, for example, a blind or sight-impaired person to (a) quickly decide whether a document is worth reading (or listening to) and (b) quickly find specific information within a paper.

Such a skimming system must extract topics, or concepts, that are important in a document and then organize them in such a way that they can be browsed quickly. Following the premise that a lexical chain can represent a concept expressed in a document [1], we use lexical chains to build browsable topic maps for scientific papers.

The domain for our skimmer is scientific articles. We have built a system for detecting lexical chains in these texts. The algorithm essentially follows the Silber and McCoy methodology [13], but uses a few modifications that are important in our text type. For instance, we have found in scientific papers that adjective modification is essential to characterize topics well.

Our users have direct contact with the lexical chains that our program outputs. We thus require a different kind of evaluation from previous work, namely one that formally evaluates the quality of the chains per se.

For example, consider the two lexical chains in Figure 1 that were automatically generated by our system[1]. Num-

ELECTRA '05 Salvador, Brazil

[1]The example paper used throughout this article is: Lee, Lillian (1999). "Measures of Distributional Similarity". *37th Annual Meeting of the ACL*, pp. 25-32.

1. confusion probability (7), probability (3), positive probability (1), conditional cooccurrence probability (1), arbitrary probability distributions (1), probability estimate (1), probability distributions (1), probability estimation (1), probabilities (3), base probabilities (1), base language model probabilities (1), verb cooccurrence probabilities (1), unigram probabilities (1), correct probabilities (1), smooth word cooccurrence probabilities (1), conditional verb cooccurrence probabilities (1)
2. values (2), appropriate values (1), actual values (1), distribution values (1)

Figure 1: Lexical chains automatically generated by our system for the example paper.

bers appearing in parentheses represent the frequency of the preceding term in the paper.

We perceive the first chain in Figure 1 as a good representation of an important topic in the paper (*probabilities*), but not the second one. We want our evaluation method, which is intrinsic, to pick up on this intuitive difference. We could choose an extrinsic evaluation of the final product (e.g. to determine if a user can solve a search task faster with our system output, than with a different document such as an abstract), but this evaluation method, like the ones in previous work, does not directly tell us to which degree lexical chains are intuitive to humans, and which chains describe the topics in a paper well. We thus opted to create a "gold standard" of lexical chains to compare against.

In this paper we report preliminary results from a small annotation study, where we asked human annotators to manually create lexical chains for two texts. We also developed and compared coverage and agreement measures that allow us to quantify the similarities between lexical chains created by different humans, and between humans and our system. We conducted these studies to shed more light on the question of how much difference there is in human intuition about lexical chains. Additionally we hope to use the human training material for a filtering component of our system, as by far the largest problem we encounter is the over-generation of lexical chains by our automatic method.

Our main contribution in this paper is the methodology of the human annotation study and the novel measures for reporting agreement between lexical chains created by different sources. We present measures for how much one single chain agrees with another single chain, and for how much a set of chains (representing one paper) agrees with another set of chains created by a different annotator.

The rest of the paper is structured as follows: the next section discusses a peculiarity of our task, namely the need to find local as well as global chains. Section 3 describes our system. The pilot study for lexical chain annotation is given in section 4. Section 5, the core of this paper, describes the coverage and agreement measures between lexical chains. We then report our system's results in terms of difference from each human annotator. The last section gives conclusions.

2. GLOBAL AND LOCAL CHAINS

A scientific paper will have main topics which describe the principal purpose(s) of the paper. It will also contain more localized topics. These may be associated with subsections

1. measures (15), distributional similarity measures (8), similarity functions (7), function (12), functions (14), divergence (15), similarity measures (8), similarity metric (3), similarity function (7), metric (8), similarity metrics (2), coefficient (7), measure (6), metrics (4)
2. distance-weighted (5), distance-weighted averaging (5), distance-weighted averaging model (1)

Figure 2: Global and local chains.

of the paper, for the purpose of providing more detailed information. Given a paper, a concept associated with a chain need not run through the entire paper ("global chain"), but can also cover only a subset of the paper ("local chain"). Schematically, this is shown in Figure 3, where the shorter chains are local, as they refer to localized topics.

A real-world example of a global chain is chain #1 in Figure 2. This chain represents the topic *similarity measures/metrics*. Terms from this chain are used throughout the example paper by L. Lee, including the title. An example of a local chain is #2 in Figure 2. This chain represents the topic *distance-weighted averaging* and is only used in section 1 (*Introduction*) and section 3 (*Empirical Comparison*.

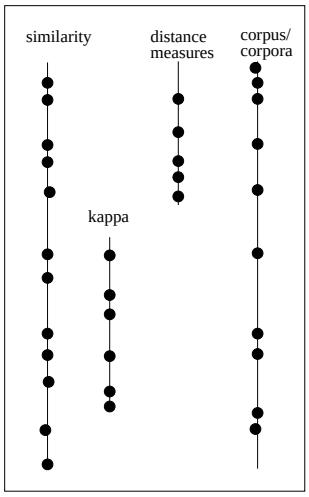

Figure 3: Examples of local and global chains

We have discussed local chains, and why they are important in our task. This motivates a slightly unorthodox approach to chain membership; we allow one term to occur in more than one chain, unlike previous definitions [1, 13]. This is necessary in order to allow for the parallel existence of global and local chains that cover similar aspects of a concept. For instance, the local chain in Figure 4 contains "divergence" as its most frequent term, a term which also belongs to a different, global chain. We would not want to have to exclude all occurrences of divergence from the global chain, just to enforce the uniqueness constraint.

3. THE SYSTEM

Our lexical chainer is based on the Silber and McCoy generation algorithm and uses the Barzilay and Elhadad scor-

divergence (15), total divergence (1), skew divergence (5), jensen-shannon divergence (7), kl divergence (5), outliers (1)

Figure 4: An example local chain.

Text	A	B	C	Average
1	3.5%	0%	1.4%	2%
2	10.2%	9.4%	0%	7%

Figure 5: Presence of WordNet relations in human-generated chains.

ing system. Both approaches used WordNet [10] to group together related words. The scoring system we use is due to Barzilay and Elhadad ([1]) scored chains independently from each other by assigning different scores to each type of relation. For instance, word repetition contributed the most points (7), followed by the synset relation (4). Each of the other relations contributed 1 point. We modify this scheme by only using synonymy and hypernymy/hyponymy, like Silber and McCoy.

We follow Silber and McCoy's ([13]) linear time algorithm for creating chains and performing word-sense disambiguation. This is done by first ranking the chains by score. If the score of the highest-scoring chain is above a set threshold, then the chain is added to the final set of lexical chains for the document. Each word in the chain is then removed from all other chains, and the chains are rescored and reranked. This is repeated until there is no chain left that has a score higher than the threshold and has not already been added to the final chain list. Each word, therefore, belongs to at most one lexical chain in the final chain list. In each approach, only nouns and noun compounds were considered for membership in lexical chains.

The differences between our lexical chainer and the Silber and McCoy system are discussed in the subsections that follow.

3.1 Multi-word scientific terms and WordNet

Scientific papers tend to use a large number of multi-word terms [7]. Such terms are usually not present in WordNet. For each noun compound which does not exist in WordNet, Barzilay and Elhadad assign the compound the WordNet synset value of the head noun. Analogous to this 'shared head' relation, we additionally define a 'shared modifier' relation. This allows two terms which share one or more modifiers to be included in the same lexical chain (e.g. *similarity measure* and *similarity distribution*. The decision of including a 'shared modifier' relation is supported by our annotation data (cf. Figure 6).

Stokes [14] found that the WordNet relations played less of a role in her lexical chains than repetition did. She attributes part of this effect to the sparcity of compound terms in WordNet. Our annotation results suggest that for scientific texts WordNet relations play an even smaller role than in news texts, as shown in Figure 5; on average, only 4% of the relations in our chains correspond to WordNet relations. This seems to be in line with Justeson and Katz's argument technical that terms tend to be repeated instead of substituted [7]. Indeed all of the WordNet relations we observed between single-word terms (e.g. *probability/chance* distributional similarity measures, similarity functions, similarity measures, similarity metric, similarity function, similarity metrics

Figure 6: An example of a lexical chain whose terms are related by premodification and require a partial overlap relation.

Text	A	B	C	Average
1	29%	42%	37%	36%
2	19%	32%	39%	30%

Figure 7: Percentage of adjectives in human-generated lexical chains.

and *data/corpus*).

Additionally we also believe that the word sense disambiguation problem is less acute in scientific text, because (1) the terms are naturally longer and thus more specific, and (2) word sense variation is lower within a subfield.

3.2 Adjectives

The importance of adjectives as premodifiers in technical terms in scientific text is well-acknowledged [7]. In addition, Justeson and Katz report that 4% of the terms in their dictionary sample are single adjectives or adjective phrases. Our approach to lexical chaining allows adjectives as modifiers in noun phrases and as heads of adjective phrases to be chain candidates.

The data produced by our annotators suggest that humans do indeed heavily incorporate adjectives into their lexical chains. Overall, 37% of all term types are adjectives or contain adjectives (30% for annotator A, 41% for annotator B, and 38% for annotator C). Of the two texts that we had annotated, Text 1 seems to have more adjectives than text 2 (cf. Figure 7). For text 1, for example, 81% of all of the human-generated chains contained at least one adjective.

Our decision to include adjectives into lexical chains is in contrast to previous work in lexical chains: Barzilay and Elhadad only allow nouns to be considered in creating terms for lexical chains, as do Morris and Hirst [11] (who work on Reader's Digest articles) and Silber and McCoy [13]. Stokes [15] uses adjectives which form part of a complex proper noun such as *Irish* in compound terms like *Irish journalist*. We believe that the importance of adjectives varies considerably with genre (Barzilay and Elhadad used newspaper texts, Morris and Hirst [12] Reader's Digest articles.

However, we believe that not all adjective in scientific text are equally important to represent a scientific text accurately. In a term such as *statistical significance*, *statistical* disambiguates the sense of *significance* whereas an adjective like *higher* in *higher significance* does not. Levi [9] calls adjectives such as *statistical* non-predicating. We have implemented an algorithm based on some of her linguistic tests to filter out predicating adjectives. This algorithm is however not the focus of the current paper; we will report about it elsewhere.

3.3 Non-uniqueness of chain membership

Silber and McCoy restrict each term to appearing in only one chain. The "best" chain for a given term is chosen and that term is removed from the rest of the chains. Our chainer

1. probability (15), probability estimation (1), conditional cooccurrences probability (2), cooccurrence probability (2), probability distribution (1), confusion probabilities (1), frequencies (3), conditional verb cooccurrence probabilities (2), relative frequencies (2), unigram probabilities (1), word cooccurrence probabilities (2), conditional probabilities (3), likelihoods (1), base probabilities (3), likely (2), probability estimate (2)
2. cooccurrences (3), cooccurrence probability (2), word cooccurrence probability (2), cooccurrence pair (1), conditional verb cooccurrence (1), verb-object cooccurrence pairs (1)

Figure 8: Membership of a term in a global and a local chain.

allows a term to appear in multiple chains. In scientific papers, a term may intuitively belong to a global topic and to a local topic. For example, chain #1 in Figure 8 represents the global topic *probability* and contains the term *cooccurrence probability*. Chain #2 in the same figure represents the local topic *cooccurrence* and also contains the term *probability cooccurrence*.

The annotation guidelines allow the possibility of using a term in more than one chain but leaves the decision up to the annotators. All three of the current annotators used at least one term in more than one chain.

4. PILOT STUDY

Measuring the extent to which human intuitions about lexical chains agree is an interesting task, both from a psycholinguistic viewpoint as well as from a practical one. A lexical chaining algorithm was first proposed by Morris and Hirst [11], based on the idea of lexical cohesion as in [4]. Even though it seems clear that most humans intuitively understand the concept of lexical chains, few experiments of the psycholinguistic plausibility of actual chain construction have been performed. Morris and Hirst [12] present a pilot study of the subjectivity of readers' perceptions of relations between words that make up lexical chains. The domain for this study was a collection of general-interest articles taken from Reader's Digest. Five subjects were asked to read the first 1.5 pages of an article and mark each word group that they perceived. For each word group, they identified pairs of related words and the relation between them. The subjects agreed on a subset of the word groups while also having individual variation.

They point out that the "degree of individual difference or subjectivity in text understanding is likely to vary with text type." It is thus necessary for us to collect annotators' perception and agreement data for the text type we work on, scientific domain.

As we already motivated, we also have practical reasons for creating a manual training set of lexical chains: we need them to directly evaluate the quality of our automatically created lexical chains, and we intend to use them as training material to learn to recognize weak chains in order to remove them from the final lexical chain set.

Because of our focus on scientific papers, we decided to also perform annotation, choosing to randomly select papers from the ACL anthology as our data.

Experimental design is as follows: We use three annotators, who are given a set of materials as described below. Annotator A is a doctoral student in computer science. Annotator B is the second author of this paper, and annotator C is the first author of this paper. The annotators are given unrestricted time to create sets of terms that they judge to be related given the context of the paper. Each set of terms then represents one lexical chain. The guidelines are four pages long and essentially describe the task as follows: A term can comprise a single word or a combination of words, all taken directly from the text. Words used in terms may be nouns, adjectives, or adverbs. Possible relationships between terms in a chain are mentioned which include inflectional variance, synonymy, hypernymy/hyponymy, holonymy, and meronymy.

There are no limits placed on the size of lexical chains or the number of chains needed to describe a document. We found that there are many intuitive similarities between chains created by our annotators. There are also many differences, such as in the number of chains used and in the exact terms that are used.

For the ongoing annotation experiment, human annotators are given a collection of materials including a list of all words in the paper together with part-of-speech tags generated by RASP [2]. Each annotator is also given a list of maximal noun phrases automatically extracted from the paper. Use of these lists is optional, but they are provided as different visualizations of the terms in the paper.

To measure the agreement between two annotators we need a metric that will do the following:

1. When comparing two lexical chains, one chain should be penalized for not covering its topic as well as a competing chain.

2. When comparing two sets of lexical chains, one chain set should be punished for not covering the paper as well as a competing chain set.

3. A chain set should be penalized for splitting chains (i.e. using two chains to describe the same topic), in comparison to having identical chains (non-split chains), but it should penalize it less than in a situation where one of the split chains is missing or replaced with irrelevant terms.

4. A chain set should be penalized for merging chains (i.e. combining multiple concepts into one chain); see above.

We use a token-based approach to comparing chains rather than a type-based approach because we believe term repetition in scientific texts to be a strong indicator of the relevance of topics.

Section 5 describes some coverage and agreement measures that we are using to evaluate lexical chains and sets of lexical chains.

5. COVERAGE AND AGREEMENT MEASURES

5.1 Comparing lexical chains

In this section we compare four measures for computing the similarity between two lexical chains. We discuss the

properties of each measure and how they affect the usefulness of the measure for our task.

When comparing two lexical chains x and y, two (not necessarily equal) agreement measurements are important:

1. The degree to which y is covered by or similar to x
2. The degree to which x is covered by or similar to y

To compute chain set agreement between two annotators (or chain set similarity between two papers), we find (for each chain in chain set A) the best match in chain set B, according to either measure detailed below. Adding together the agreement scores for each match gives us Equation 1.

$$m(A,B) = \sum_{x \in A} \frac{m_1(x,B)|x|}{|A|}. \quad (1)$$

$m(A,B)$ measures the degree to which all chains in A cover any of the chains in B.

5.2 Cosine measure

For a baseline, we use the standard cosine metric. Each lexical chain is represented as a vector of term frequencies. Of the measures considered here, the cosine metric is the only one that is symmetric.

5.3 KL distance

Another comparison measure that we evaluated is the Kullback-Leibler (KL) distance [8]. It is a measure of similarity between two distributions, as defined in Equation 2.

$$KL(P,Q) = \sum_{i=0}^{n} p_i \log_2(\frac{p_i}{q_i}), \quad (2)$$

where $P = (p_0, ..., p_n)$ and $Q = (q_0, ..., q_n)$ are probability distributions.

We compare chains by representing each chain as a vector of relative term frequencies. Suppose we wish to compare chains X and Y. Since both distributions in Equation 2 must contain the same number of points, we set the length of the vector for chain X and the length of the vector for chain Y equal to the order of the union of the terms in X and Y. This means that for two chains that do not have exactly the same terms, their corresponding vectors will contain 0-values representing terms missing from the chain. Since each value in P and Q must be nonzero for KL, we perform simple add-one smoothing.

5.4 Strict term overlap

We also consider simple term overlap

$$c(x,y) = \frac{|x \cap y|}{|y|}. \quad (3)$$

Two chains x and y are treated as sets of tokens (with multiplicity).

We measure the coverage of B by x as

$$m_1(x,B) = \max_{y \in B} c(x,y). \quad (4)$$

Similarly, we measure the coverage of x by B as

Measure	A→B	A→C	B→C
Cosine	18%	0%	7%
KL	82%	62%	71%
Strict overlap	100%	62%	71%
Partial overlap	100%	69%	86%

Figure 9: Agreement between automatically matched chains and manually matched chains for Text 1.

Measure	B→A	C→A	C→B
Cosine	8%	0%	8%
KL	38%	42%	71%
Strict overlap	69%	84%	67%
Partial overlap	69%	84%	67%

Figure 10: Agreement between automatically matched chains and manually matched chains for Text 1.

$$m_2(x,B) = \max_{y \in B} c(y,x). \quad (5)$$

Note that $m_1(x,B)$ and $m_2(x,B)$ need not be maximized by the same $y \in B$.

5.5 Partial term overlap

We modify our overlap measure by allowing partially overlapping terms to count as partial matches. The overlap measure in Equation 3 only recognizes exact term matches, but semantics is shared between terms even if there is a partial overlap (e.g., in modifiers or heads). We assign a weight of 0.3 to this relation.

5.6 Preliminary results

5.6.1 *Testing the measures*

To test the four measures described above, we look for the strongest chain matches between two annotators. That is, given two annotators A and B, each chain from annotator A is matched with the most similar chain from annotator B, and vice versa. Performing this task using each measure gives us four sets of chain matches for each annotator pair (going one direction). Each set of matches is then compared to a manually generated set of chain matches for the same annotator pair.

As we can see in Figures 9-13, the cosine metric performs badly when matching chains. This is primarily because a metric based on the inner product of two vectors does not issue a penalty when vectors of different lengths are compared (an attractive property in IR when comparing documents with queries). Thus, the chains that are found to match using this metric may have high frequency terms in common but may also contain several other terms not shared by other chains.

The shared-modifier algorithm had a slight improvement over the overlap measure when finding chain matches, and thus outperformed the KL distance and the baseline.

5.6.2 *Comparing rankings*

Using the chain match scores given by the measures, we can rank the strength of the chain matches. For each mea-

Measure	A→B	A→C	C→B
Cosine	29%	17%	27%
KL	86&	67%	73%
Strict overlap	86%	83%	64%
Partial overlap	86%	83%	91%

Figure 11: Agreement between automatically matched chains and manually matched chains for Text 2.

Measure	B→A	B→C	C→A
Cosine	0%	30%	13%
KL	88%	80%	75%
Strict overlap	75%	80%	75%
Partial overlap	100%	80%	75%

Figure 12: Agreement between automatically matched chains and manually matched chains for Text 2.

sure we compare the top five chains to a manual ranking of the top five chain matches for each annotator pair. We only consider the top five matches to avoid having to compare match strengths between chains with little in common. Since the cosine metric performed so poorly when finding matches, we only compared rankings for the other three measures.

Figure 14 shows the agreement between top matches selected by the different measures and those selected manually.

6. CONCLUSIONS AND FUTURE WORK

Our main contributions in this paper are our methodology of the human annotation study and a comparison of four similarity measures (including a new measures based on shared modifiers) for reporting agreement between lexical chains created by different sources. Our annotation study covers the scientific domain with the goal of training a lexical chaining system for scientific papers.

This pilot study explores the extent to which human-generated lexical chains agree in the domain of scientific texts. In future work, we will investigate the role that non-uniqueness of term membership plays in creating local chains. As we build our gold standard we hope to determine the importance of adjectives in human-generated lexical chains in the scientific domain.

Limitations of our preliminary study are:

1. We have too few annotators and use too few papers for an extensive study of lexical chain agreement in the scientific domain. This will be expanded in later work.

2. Our coverage and agreement measures do not yet handle all of the cases that we want to consider (e.g. the merger of two chains). The comparison ranking produced by our measures and presented in this paper compares well with an intuitive ranking for the most important matches, but compares badly overall.

Measure	Text 1	Text 2
Cosine	6%	20%
KL	61%	78%
Strict overlap	74%	76%
Partial overlap	78%	86%

Figure 13: Average agreement between automatically matched chains and manually matched chains.

Match	KL	Strict overlap	Partial overlap
A→B	2	3	3
B→A	2	2	2
B→C	2	2	2
C→B	3	2	2
A→C	2	4	4
C→A	4	2	2

Figure 14: Number of chain matches ranking in the top five as compared to the manually ranked top five matches.

Future work will address the problems mentioned above.

7. REFERENCES

[1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*. ACL Madrid, 1997.

[2] E. Briscoe and J. Carroll. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas, Canary Islands, May 2002. (LREC 2002).

[3] S. Green. Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering*, 1999.

[4] M. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[5] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 1997.

[6] G. Hirst and St-Onge. Lexical chains as representation of context for the detection and correction of malapropisms. In *Fellbaum, C., ed., WordNet: An Electronic Lexical Database and Some of its Applications*. The MIT Press, 1998.

[7] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

[8] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statsistics*, 22:79–86, 1951.

[9] J. Levi. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, 1978.

[10] R. F. C. G. D. Miller, G.; Beckwith and K. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4).

[11] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 1991.

[12] J. Morris and G. Hirst. The subjectivity of lexical cohesion in text. In *Shanahan, James G.; Qu, Yan;*

1. **similarity**, measures (15), distributional similarity measures (8), similarity functions (7), function (12), functions (14), divergence (15), similarity measures (8), similarity metric (3), similarity function (7), metric (8), similarity metrics (2), coefficient (7), measure (6), metrics (4)
2. **support**, **regions**, supports (7), regions of positive probability (1), support-intersection data (1), support (3), support regions (1)
3. **similarity**, distributional similarity (7), similar (7), distance-weighted (5), distributional similarity (7), semantic similarity (1), distance (2), dissimilarity (1), similarity-based (1), similarity (28), commonality (2), differences (3)
4. **unseen**, unseen cooccurrences (2), sparse data (2), low frequency events (1), unseen events (2), unseen word pair (1), unseen (8), unseen pairs (1), sparseness (1)
5. **cooccurrence**, **cooccurrences**, cooccurrence (6), cooccurrences (3), word cooccurrence (2), neighborhood (1), closest neighbors (1), nearest neighbors (2)
6. **probability**, **estimate**, **estimation**, probability estimation (1), estimate (4), similarity-based estimation (1), probability estimate (2), estimates (2)
7. **comparison**, **empirical**, empirical comparison (3), comparison (5)
8. **distributions**, potential proxy distributions (2), distributional (8), probability distributions (2), distributions (8), potential proxy distributions (2), joint distribution (1), product distribution (1)
9. **training**, **corpus**, training corpus (2), training set (2), training partition (1), training corpus (2), training data (4)
10. **probabilities**, probability (15), conditional cooccurrence probability (2), probability distributions (2), chance (2), distributions (8), probabilities (10), verb cooccurrence probabilities (2), smooth word cooccurrence probabilities (2), base language model probabilities (2), confusion probability (7), unigram probabilities (1), likelihoods (1), conditional verb cooccurrence probabilities (1), mathematical certainty (1)
11. **events**, **data**, events (5), data (11), bigrams (1), words (7), word pair (1), cooccurrences (3), words (7), data (11), nouns (6), verbs (11), cooccurrence pair (1), noun (2), corpus (2), adjectives (1), pairs (4), noun-verb pair (1), noun-verb-verb triple (1), test triple tokens (1), test instance (1)
12. **probability**, **probability distributions**, probability (15), probability distributions (2), chance (2), average (6), statistically (2), insignificant (1), unsmoothed (1), frequencies (3), smooth (2), relative frequencies (2), likelihoods (1), joint distribution (1), product distribution (1), unigram frequencies (1), error rate (4), statistic (1), t-test (1), significance level (1), mathematical certainty (1)

Figure 15: Annotator A's lexical chains for the example paper.

and Wiebe, Janyce (Eds.) Computing attitude and affect in text, Dordrecht, The Netherlands, 2005.

[13] H. G. Silber and K. F. McCoy. Efficiently computed lexical chains as an intermediate representation in automatic text summarization. *Computational Linguistics*, 28(4), 2002.

[14] N. Stokes. Spoken and written news story segmentation using lexical chaining. In *Proceedings of the Student Workshop at HLT-NAACL 2003, Companion Volume*, pages 49–54, Edmonton, Canada, 2003.

[15] N. Stokes. *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain*. Department of Computer Science, University College Dublin, 2004.

Appendix

Figures 15 and 16 show all of the chains constructed by two of the annotators for the example paper. Numbers appearing in parentheses represent the frequency of the preceding term in the paper. Terms appearing in bold are chain representatives and were automatically extracted from the manual chains.

1. **similarity**, similarity (28), distributional similarity measures (8), similarity functions (7), distributional similarity (7), semantic similarity (1), similarity measures (8), similarity metric (3), similar words (5), new similarity metrics (2), extreme dissimilarity (1), similarity-based estimation (1), inherently better similarity ranking (2), good similarity metric (3)
2. **probabilities**, probability (15), probability estimation (1), conditional cooccurrences probability (2), cooccurrence probability (2), probability distribution (1), confusion probabilities (1), frequencies (3), conditional verb cooccurrence probabilities (2), relative frequencies (2), unigram probabilities (1), word cooccurrence probabilities (2), conditional probabilities (3), likelihoods (1), base probabilities (3), likely (2), probability estimate (2)
3. **distribution**, distribution (6), proxy distributions (2), probability distribution (1), average distribution (1), joint distribution (1), product distribution (1), empirical distribution (1)
4. **unseen**, sparse data (2), low frequency events (1), unseen cooccurrences (2), unseen word pair (1), unseen (8), unseen pairs (1), sparseness (1)
5. **training**, training set (2), estimate (4), training partition (1), test-set bigrams (1), training corpus (2), test sets (1), test-set performance (2)
6. **concurrence**, cooccurrences (3), cooccurrence probability (2), word cooccurrence probability (2), cooccurrence pair (1), conditional verb cooccurrence (1), verb-object cooccurrence pairs (1)
7. **method**, **backoff**, backoff method (2), interpolation method (1), backoff smoothing method (1)
8. **distance-weighted**, **averaging**, distance-weighted (5), distance-weighted averaging (5), distance-weighted averaging model (1)
9. **divergence**, divergence (15), total divergence (1), skew divergence (5), jensen-shannon divergence (7), kl divergence (5), outliers (1)
10. **significant**, statistically significant (2), significant (3), significance level (1)
11. **evaluation**, **pseudoword disambiguation task**, evaluation (3), pseudoword disambiguation task (1), empirical results (1), decision task (3), empirical comparison (3), evaluation methodology (1), binary decision task (3), experimental framework (1), correct answer (1), paired t-test (1), prediction tasks (1)
12. **information theoretic metric**, **similarity metric**, information-theoretic metric (1), similarity metric (3), similarity measures (8), cosine metric (2), jaccard coefficient (1), jensen-shannon divergence (7), kl divergence (5), nonparametric measure (1), correlation (1), mutual information (1), value difference metric (2), dice coefficient (1), l2 norm (1), l1 norm (3), statistic (1), euclidean distance (1), skew divergence (5), alpha - skew divergence (5), good similarity metric (3), similarity function schema (1)
13. **performance**, **average**, performance (10), precision (1), average performance (3), average error rate (4), test-set performance (2)
14. **nouns**, **verbs**, nouns (6), verbs (11), transitive verbs (1), head noun (1), direct object (1), similar adjectives (1), frequent nouns (1), noun-verb pair (1), noun-verb-verb triple (1)
15. **smoothing**, **unsmoothed**, smoothing (1), unsmoothed (1), smoothed base language model (5)
16. **model**, **language**, language model (5), language model probabilities (2), language modeling (1), smoothed base language model (5), model (9)
17. **neighbors**, neighborhood (1), neighbors (3), nearest neighbors (2)
18. **function**, **weighting**, **weight**, weighting (1), weight function (1)
19. **substitutability**, substitutability (1)
20. **generalization**, **asymmetric**, **novel**, **symmetric**, symmetric (2),s novel asymmetric generalization (1)
21. **information**, **negative**, negative information (1)
22. **translations**, **mutual**, translations (1), mutual translations (1)

Figure 16: Annotator B's lexical chains for the example paper.