Introduction to Computational Semantics and its Applications

Ekaterina Kochmar
Computer Laboratory, University of Cambridge
Automated Language Teaching and Assessment (ALTA) Institute

UCL, March 2018
Computational Linguistics
Introduction to Computational Linguistics: a bit of history

• Computational Linguistics originated in the U.S. in 1950s
• Focused on Machine Translation, particularly from Russian to English
• Deemed to be an easy computational task
• Note: this task is not perfectly solved even today...
Computational Linguistics and other fields

Theoretical Linguistics
- fundamental questions; theoretical basis

Computational Linguistics
- computational approaches to linguistic questions

Natural Language Processing
- applications; processing of large amounts of data
Computational Linguistics and other fields

Theoretical Linguistics

Computational Linguistics

Natural Language Processing

Machine Learning

helps to learn from data & detect regularities

Artificial Intelligence

language understanding & language generation
## Computational Linguistics vs Theoretical Linguistics

<table>
<thead>
<tr>
<th>Theoretical Linguistics</th>
<th>Computational Linguistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>✦ develops linguistic theory</td>
<td>✦ builds computational models</td>
</tr>
<tr>
<td>✦ seeks to answer fundamental questions</td>
<td>✦ seeks to confirm and test fundamental approaches</td>
</tr>
<tr>
<td>✦ is based on theoretical approaches</td>
<td>✦ rule-based or statistical, data-driven approaches</td>
</tr>
<tr>
<td>✦ theory-oriented</td>
<td>✦ application-oriented</td>
</tr>
</tbody>
</table>
Computational & Theoretical Linguistics: Fields & Tasks

- Phonology/phonetics → speech processing, speech recognition
- Morphology → morphological analysis, stemming, lemmatisation
- Word level: word segmentation, part-of-speech tagging, language modelling
- Syntax → parsing
- Semantics → lexical and computational
- Discourse and pragmatics → discourse analysis
• **Speech analysis:** based on what we know about phonetics and phonology, can we recognise speech, i.e. transcribe the audio signal as text?

• **Speech synthesis:** Can we generate the speech signal based on text?

* Here and on the other slides: the images are adopted from Jurafsky and Martin. *Speech and Language Processing*. Second edition. 2009
• What is a basic linguistic unit? → **Word**?
  • Is ‘**U.S.**’ one word?
  • Is ‘**theory-based**’ one word?
  • Is ‘.’ part of the word as in ‘**Mr.**’?
  • What about ‘;(;)?

• The notion of a word depends on language:
  • **FR** “l’ensemble”
  • **GER** “**Lebensversicherungsgesellschaftsangestellter**” = Lebens-versicherungs-gesellschafts-angestellter = ‘life insurance company employee’

* Here and on the other slides: the mages are adopted from Jurafsky and Martin. *Speech and Language Processing*. Second edition. 2009
Fields & Tasks:
Text segmentation & normalisation

• Chinese – no spaces to separate words
  • 莎拉波娃现在居住在美国东南部的佛罗里达。
  • 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  • Sharapova now lives in US southeastern Florida

• Japanese – many alphabets mixed

フォーチュン500社は情報不足のため 時間あたり$500K(約6,000万円)

Katakana  Hiragana  Kanji  Romaji

* Here and on the other slides: the mages are adopted from Jurafsky and Martin. Speech and Language Processing. Second edition. 2009
• Words are built of smaller units – morphemes

• Morphology: *inflectional* (to express grammatical category) and *derivational* (to change the lexical category in related words)

• Richness of plural form morphology in English:
  • word → word\text{\text{s}}, book → book\text{\text{s}}
  • fox → fox\text{\text{es}}, hero → hero\text{\text{es}}
  • ax → ax\text{\text{es}} and axe\text{\text{s}} ← axe
  • city → citi\text{\text{es}}, morphology → morphologi\text{\text{es}}
  • leaf → leav\text{\text{es}}, shelf → shelv\text{\text{es}}
  • foot → feet, man → men, mouse → mice
  • corpus → corpora, phenomenon → phenomena
Richness of morphological forms in many other languages is higher:

- cf. Turkish: *Uygarlastiramadiklarimizdanmissinizcasina* –
  ‘(behaving) as if you are among those whom we could not civilise’ =
  *Uygar - las - tir - ama - dik - lar - imiz - dan - mis -siniz-casina* =
  ‘civilised’-‘become’-‘cause’-‘not able’-‘past’-‘plural’-‘p1pl’-‘abl’-‘past’-‘2pl’-‘as if’

With the computational models we want to recognise:

- *book* and *books* – {book}; *is, are, was, been* – {be} --> **lemmatisation**
- *automate, automation, automated, automatic* – {automat} --> **stemming**

* Here and on the other slides: the mages are adopted from Jurafsky and Martin. Speech and Language Processing. Second edition. 2009
Fields & Tasks:

Sequence labelling and modelling

- **Part-of-speech tagging**
  - We can fish vs We can fish
    - PRON AUX VB
    - PRON VBP NOUN

- **Language modelling:**
  - lectu__
  - Today’s lecture will take ___
* Here and on the other slides: the images are adopted from Jurafsky and Martin. Speech and Language Processing. Second edition. 2009
**Fields & Tasks:**

**Semantics**

- **Lexical Semantics:** word senses and relations between word senses
  - *I went to the bank and withdrew money from my account*
  - *I went to the bank and had a walk along the river*

- **Computational Semantics (Vector Semantics):** representation of word (and larger linguistic units) meaning in a shared semantic space

*Here and on the other slides: the images are adopted from Jurafsky and Martin. Speech and Language Processing. Second edition. 2009*
Computational Semantics
Our goal is to build a computational model of word meaning so that a machine can understand the words, derive the meaning of phrases and detect the anomalies.

Luckily, there are compositional distributional (as well as distributed) semantic models that can help us:

- **distributional/distributed models** helps capturing individual words’ meaning.

- **compositional semantic models** help successfully (or unsuccessfully) combine the individual meanings into the meaning of a longer phrase.
Mikolov et al. (2013) showed that computers can reason about word meaning similarly to humans using an example of word analogy:

$\text{Man}$ is to $\text{woman}$ as $\text{king}$ is to $\text{____}$?
Mikolov et al. (2013) showed that computers can reason about word meaning similarly to humans using an example of word analogy:

\[ \text{Man is to woman as} \]
\[ \text{king is to queen?} \]

What the solution boils down to is:

\[
\text{MEANING(WORD) = MEANING(king) - MEANING(man) + MEANING(woman)}
\]
How do we know what words mean?

Who is a queen?
Computational Semantics: Learning through experience

- We read about kings and queens
- We hear about them on the news
- We see them on the TV or, perhaps, even in person
- => We build our semantic model of what the words king and queen mean based on our experience
- How can a machine learn the meaning of a word?
Computational Semantics: Key assumptions of distributional semantics

- **Key assumption**: word meaning can be approximated by a word’s distribution

  “You shall know a word by the company it keeps” (Firth)

- **Method**: represent words with distributional vectors, dimensions = co-occurrence with a predefined set of context words

- **Hypothesis**: semantically similar words occur in similar contexts and, therefore, will be represented with similar vectors in the semantic space

- A nice property of a direct interpretation of word meaning through vectors in space
Computational Semantics: Word distributions

Her Majesty the Queen
The Queen's speech during the State Visit to...
Buckingham Palace is the Queen's official London residence...
The Crown of Queen Elizabeth
The Queen Mother
### Computational Semantics: Word vectors

<table>
<thead>
<tr>
<th></th>
<th>he</th>
<th>she</th>
<th>royal</th>
</tr>
</thead>
<tbody>
<tr>
<td>queen</td>
<td>20</td>
<td>581</td>
<td>389</td>
</tr>
<tr>
<td>king</td>
<td>599</td>
<td>18</td>
<td>344</td>
</tr>
</tbody>
</table>
Computational Semantics: Distributional Semantic Models

Represent words as vectors

How should we build them?

What are the dimensions?

Learn from the data

Build vectors using the surrounding words

$\rightarrow$ Distributional models of word meaning
Computational Semantics: Word meaning representations

- **Distributional models**: build word vectors using contexts

- **Distributed models** (word embeddings): dense low-dimensional (300) representations where each dimension encodes some distinct property
Computational Semantics: Word meaning representations

- **Distributional models**: build word vectors using contexts

- **Distributed models** (word embeddings): dense low-dimensional (300) representations where each dimension encodes some distinct property

- Essentially: different ways to build **word vectors**

- A bit of math:
  - How to measure semantic similarity? Use cosine (distance) measure

\[
\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]
Mikolov et al. (2013) showed that computers can reason about word meaning similarly to humans using an example of word analogy:

\[
\text{Man is to woman as } \quad \text{king is to queen?}
\]

What the solution boils down to is:

\[
\text{REP(WORD)} = \text{REP(king)} - \text{REP(man)} + \text{REP(woman)}
\]
Check your intuitions

Input: *Russia* is to *Moscow* as *China* is to ___?

- France
- Germany
- Greece
- Italy
- Japan
- Poland
- Portugal
- Spain
- Turkey
Computational Semantics & Second Language Learning
Learner Errors

English Today

- About 7,000 known living languages
- Native speakers of English – about 5.52%
- The rest – non-native speakers (language learners)
- The University of Cambridge: 18,000 students, of which 3,500 are international students from >120 different countries
Learner Errors
Why this matters

In scientific text, it is particularly important that the ideas are clearly expressed.

What we aim to do:
- analyse the text
- detect the problematic areas
- suggest corrections
- ideally, do all of the above automatically

Keywords: Text classification, hierarchical classification, feature selection, feature weighting

Abstract. In recent years, there have been extensive studies and rapid progresses in automatic text classification, which is one of the hotspots and key techniques in the information retrieval and data mining field. Feature extraction and classification algorithm are the crucial technologies for this problem. This paper firstly proposed feature extraction algorithm based on key words, the algorithm selected key words set from special part of scientific papers, and employed mutual information to extract features. And then, proposed an improved hierarchical classification method, and realized hierarchical classification of Chinese scientific papers.

Introduction

Goal of automatic text classification system is an orderly organization of the text sets, to organize the similar and related texts together. As a tool of knowledge organization, it provides more effective search strategies and more accurate query results for information retrieval.[1]
Learner Errors
State-of-the-art

• Currently, widely used spell-checkers and grammar-checkers can only detect and correct a limited set of errors (e.g., spelling, typos, some grammar)

• However, if you’ve picked a completely incorrect word they are unlikely to ask you if you have “meant powerful computer instead of strong computer?”
Does incorrect word choice impede understanding?

<table>
<thead>
<tr>
<th>Error</th>
<th>Correction</th>
<th>Error type</th>
<th>Problematic to understand?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am * student</td>
<td>I am a student</td>
<td>Missing article</td>
<td></td>
</tr>
<tr>
<td>Last year I went <em>in</em> London on a business trip</td>
<td>Last year I went to London on a business trip</td>
<td>Wrong preposition chosen</td>
<td></td>
</tr>
<tr>
<td><em>big</em> history <em>large</em> knowledge ...</td>
<td>long history broad knowledge ...</td>
<td>Wrong adjective chosen</td>
<td></td>
</tr>
</tbody>
</table>
## Learner Errors Issues

Does incorrect word choice impede understanding?

<table>
<thead>
<tr>
<th>Error</th>
<th>Correction</th>
<th>Error type</th>
<th>Problematic to understand?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am * student</td>
<td>I am a student</td>
<td>Missing article</td>
<td>?</td>
</tr>
<tr>
<td>Last year I went <em>in</em> London on a business trip</td>
<td>Last year I went to London on a business trip</td>
<td>Wrong preposition chosen</td>
<td>?</td>
</tr>
<tr>
<td><em>big</em> history <em>large</em> knowledge ...</td>
<td>long history broad knowledge ...</td>
<td>Wrong adjective chosen</td>
<td>✓</td>
</tr>
</tbody>
</table>
Depending on the word type, the change in the original meaning can be significant:

When somebody uses an expression *big history* do they mean “academic discipline which examines history from the Big Bang to the present”?
Back to linguistics...

<table>
<thead>
<tr>
<th>Function words</th>
<th>Content words</th>
</tr>
</thead>
<tbody>
<tr>
<td>✦ link and relate the words to each other</td>
<td>✦ express the meaning of the expression</td>
</tr>
<tr>
<td>✦ are very frequent in language</td>
<td>✦ are conceptual units</td>
</tr>
<tr>
<td>✦ examples – articles and prepositions:</td>
<td>✦ examples – nouns, verbs and adjectives:</td>
</tr>
<tr>
<td></td>
<td>I am a student</td>
</tr>
<tr>
<td></td>
<td>at the University of Cambridge</td>
</tr>
<tr>
<td></td>
<td>I study Computer Science at the University of Cambridge. The course is very intensive</td>
</tr>
</tbody>
</table>
Content Words
How to solve the task of ED in content words?

• Errors in content words (nouns, verbs, adjectives) are diverse → difficult to generalise and learn regularities from the data

• The contexts are also more diverse → we might never see exactly the same context around content words again and learn anything about the features

• Corrections cannot be represented as a finite set applicable to all nouns, all verbs or all adjectives in language, and they always depend on the original incorrect word

• Content words are not just linking other words, they express meaning → we should take semantics into account
Content Words
Types of errors in content words

• Words are confused because they are **similar in meaning**: 

  He gave a *small speech* *(short speech)*

• Words are confused because they have **similar form**: 

  *It includes articles over ancient Greek sightseeings as the Alcropolis or other famous places* *(ancient sites)*

• There are some other, **less obvious** reasons: 

  *Deep regards, John Smith* *(kind regards)*

• Interpretation depends on the **context**, and the chosen words simply don’t fit: 

  *The company had great turnover, which was noticable in this market* *(high turnover)*
## Semantic Approach

### Semantic Space construction

<table>
<thead>
<tr>
<th></th>
<th>give</th>
<th>last (v)</th>
<th>build</th>
<th>topic</th>
<th>big</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>speech</strong></td>
<td>85</td>
<td>18</td>
<td>0</td>
<td>33</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td><strong>talk</strong></td>
<td>84</td>
<td>23</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td><strong>house</strong></td>
<td>0</td>
<td>2</td>
<td>67</td>
<td>0</td>
<td>56</td>
<td>...</td>
</tr>
</tbody>
</table>
Semantic Approach
Can any language expression be modeled this way?

What happens when we try applying same models to longer expressions?

- We might find 100 examples with the word *speech*, 50 of which will be about *long speech*, 2 about *45-minutes speech* and none about *7-minutes speech* (or *small speech*)

- That means, longer expressions (*1-hour speech*, *1-hour long speech*) will necessarily have sparser and less reliable vectors

- Also, we won’t be able to say anything about either *7-minutes speech* or *small speech* – if we don’t see it in the data, does it means both are implausible / nonsensical? Have we just not looked carefully enough?
Semantic Approach
Compositional Semantics methods

Instead of relying on distributional information for longer phrases, let’s use distributions of words within phrases and build vectors for longer phrases in a compositional way.

- **Component-wise additive** model:
  \[ c_i = a_i + b_i \]
  
  
  \[(\text{small} \_ \text{speech})_i = \text{small}_i + \text{speech}_i \]

- **Component-wise multiplicative** model:
  \[ c_i = a_i \times b_i \]
  
  \[(\text{small} \_ \text{speech})_i = \text{small}_i \times \text{speech}_i \]
Semantic Approach
Measures of semantic anomaly

• Earlier, we have assumed that the computational semantic representation of words will tell us something about correctness of our examples.

• Once we have modeled the phrases computationally, how can we distinguish between the representations for the correct and for the incorrect phrases?

• Since there is a direct geometric interpretation for the semantic vectors, we assume that certain properties of the vectors will highlight the differences.
Semantic Approach
Vector length as a measure of semantic anomaly

In anomalous combinations, the counts in the input vectors are distributed differently → some “incompatible dimensions” would receive low counts → anomalous phrase vectors are expected to be shorter than vectors of the acceptable phrases.

\[ \| X \|_2 := \sqrt{x_1^2 + \cdots + x_n^2}. \]

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>speech</th>
<th>small</th>
<th>short + speech</th>
<th>short × speech</th>
<th>small + speech</th>
<th>small × speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>88</td>
<td>92</td>
<td>0</td>
<td>180</td>
<td>8096</td>
<td>92</td>
<td>0</td>
</tr>
<tr>
<td>speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small</td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td></td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td>short + speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short × speech</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small + speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>small × speech</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{len}(\text{short + speech}) = 180 \quad \text{len}(\text{short × speech}) = 8096 \]
\[ \text{len}(\text{*small + speech}) = 97 \quad \text{len}(\text{*small × speech}) = 60 \]
Semantic Approach
Cosine to the input words as a measure of semantic anomaly

Anomalous phrases are less similar to the input nouns (verbs, adjectives), and the semantic space provides a direct interpretation of the similarity of two words via their distance in the space → vectors of the anomalous word combinations are expected to have **lower cosine (similarity)** to the input noun/verb/adjective vectors.
Semantic Approach

Neighbourhood density as a measure of semantic anomaly

Anomalous phrase vectors are expected to not have any specific meaning → they are expected to not be closely surrounded by other words with similar meaning → have sparser neighbourhoods in the semantic space. We measure this as an **average cosine** (= distance) to the 10 nearest neighbours.
Semantic Approach
Component overlap as a measure of semantic anomaly

We assume semantically acceptable phrases to be placed in the neighbourhoods populated by similar words and combinations, and calculate the proportion of neighbours containing the same words as the input phrases. We expect this proportion to be lower for the anomalous phrases (lower overlap)

<table>
<thead>
<tr>
<th>short speech</th>
<th>small speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>[x] speech</td>
<td>quantity</td>
</tr>
<tr>
<td>short [x]</td>
<td>small amount</td>
</tr>
<tr>
<td>talk</td>
<td>person</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Semantic approach: Machine Learning classifier for ED

- We apply **Decision Tree Classifier** to our task
- Two classes – **correct** (0) and **incorrect** (1)
- At each node, the classifier checks whether the value of the feature falls within a certain value interval (e.g., whether \( \text{len}<0.5 \) or \( \text{len} \geq 0.5 \)) and follows the relevant path
- The algorithm makes sure the most discriminative rules are applied first
### Semantic approach: Results

<table>
<thead>
<tr>
<th>Content word combinations</th>
<th>Accuracy (averaged over 5 folds)</th>
<th>Lower bound (=majority class distribution)</th>
<th>Upper bound (=annotator agreement)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective-noun</td>
<td>0.6535 ± 0.0189</td>
<td>0.5084</td>
<td>0.7467 ± 0.0221</td>
</tr>
<tr>
<td>verb-noun</td>
<td>0.6491 ± 0.0188</td>
<td>0.6086</td>
<td>0.8467 ± 0.0377</td>
</tr>
</tbody>
</table>
ED System
Further evaluation of the ED system

- **Precision** = \(#(\text{instances that belong to class } n \text{ & are identified by the system as belonging to class } n) \) / \(\#(\text{all instances identified by the system as belonging to class } n)\)

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

- **Recall** = \(#(\text{instances that belong to class } n \text{ & are identified by the system as belonging to class } n) \) / \(\#(\text{instances in the data that actually belong to class } n)\)

\[
\text{Recall} = \frac{tp}{tp + fn}
\]

- **F-measure** – harmonic mean of the two

\[
F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

<table>
<thead>
<tr>
<th></th>
<th>Predicted (+)</th>
<th>Predicted (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual (+)</td>
<td>(tp)</td>
<td>(fn)</td>
</tr>
<tr>
<td>Actual (-)</td>
<td>(fp)</td>
<td>(tn)</td>
</tr>
</tbody>
</table>
## ED System

### Class-specific performance of the ED system

<table>
<thead>
<tr>
<th>Content word combinations</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective-noun, correct</td>
<td>0.6173</td>
<td>0.7226</td>
<td>0.6558</td>
</tr>
<tr>
<td>adjective-noun, incorrect</td>
<td>0.7071</td>
<td>0.5898</td>
<td>0.6409</td>
</tr>
</tbody>
</table>
**ED System**

Class-specific performance of the ED system

<table>
<thead>
<tr>
<th>Content word combinations</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb-noun, <strong>correct</strong></td>
<td>0.6027</td>
<td>0.3192</td>
<td>0.4174</td>
</tr>
<tr>
<td>verb-noun, <strong>incorrect</strong></td>
<td>0.6637</td>
<td>0.8630</td>
<td>0.7503</td>
</tr>
</tbody>
</table>
ED System
Summary on the ED system

• We have showed that our algorithm detects errors with high accuracy

• There is still some room for improvement – it is close to, but does not yet reach human performance on this task

• The features derived using semantics and trying to capture the meaning of the words are useful

• The algorithm shows high precision → it is reliable → learners can use it to detect errors in their writing
Thank you!

• Further information:
  • http://www.cl.cam.ac.uk/~ek358/
  • Ekaterina.Kochmar@cl.cam.ac.uk

• Datasets:
  • http://www.cambridgeenglish.org
  • http://www.cl.cam.ac.uk/~ek358/an-dataset.xml
  • http://ilexir.co.uk/applications/adjective-noun-dataset/

• Useful resources: