Automated Acquisition of Multiword Expressions for Robust Deep Parsing

Valia Kordoni

Dept. of English, Humboldt-Universität zu Berlin, Germany

PARSEME 2nd General Meeting
10-11 March 2014, Athens, Greece
Program of WG2

Outline

Outline

Background

Multiword Expressions

- Syntactic or semantic properties cannot be derived from their parts [Sag et al., 2002, Villavicencio, 2005]
- phrasal verbs (e.g. *come along*), nominal compounds (e.g. *frying pan*), institutionalised phrases (e.g. *bread and butter*)
- equivalent in number to single words in speakers' lexicon [Jackendoff, 1997]
- fixed (*ad hoc*) vs flexible (*touch/find a nerve*) expressions
- opaque (*kick the bucket*) vs transparent (*eat up*) semantics

Motivation

Challenge for NLP

It is difficult to provide a unified account for the detection of these distinct but related phenomena.

Grammar Engineering and Robust Deep Parsing

- Lexical coverage is the major barrier to broad-coverage linguistically deep processing
- MWEs comprise a significant part of the missing lexicon

Motivation

Challenge for NLP

It is difficult to provide a unified account for the detection of these distinct but related phenomena.

Grammar Engineering and Robust Deep Parsing

- Lexical coverage is the major barrier to broad-coverage linguistically deep processing
- MWEs comprise a significant part of the missing lexicon

Outline

Error Mining [van Noord, 2004]

Parsability

$$R(w_i \ldots w_j) = \frac{C(w_i \ldots w_j, OK)}{C(w_i \ldots w_j)}$$

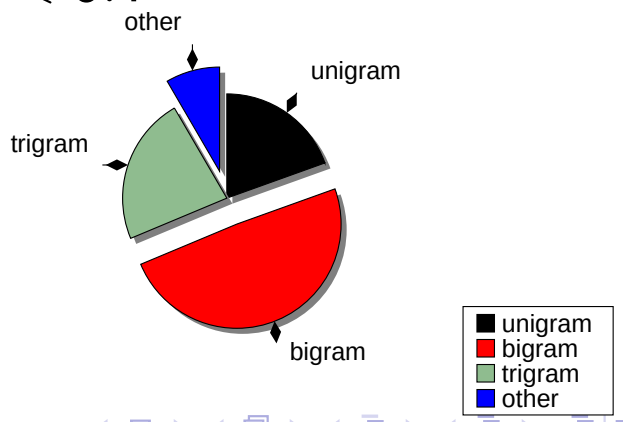
- If the parsability of a particular word sequence is very low, it indicates that something is wrong
- Parsabilities can be calculated efficiently for large corpora with suffix arrays and perfect hashing [Lucchesi and Kowaltowski, 1993]

Error Mining Experiment

- Experiment was run on BNC: the parsed sentences and the unparsed sentences (with full lex. span)
- Low parsability n-grams were extracted
- 3+ grams were taken for further inverstigation

	Num.	%
uni-gram	798	20.84%
bi-gram	2,011	52.52%
tri-gram	937	24.47%

Table : Distribution of N-grams with $R < 0.1$



Example of Low Parsability N-grams

N-gram	R	Count
the burden of	0.000	49
by and large	0.000	37
face of it	0.000	34
frame of mind	0.000	23
points of view	0.000	20
hair and a	0.000	17
the to infinitive	0.000	15
of alcohol and	0.000	8
a great many	0.083	44
glance up at	0.083	33
for and against	0.086	21
from of government	0.142	6

Example of Low Parsability N-grams

N-gram	R	Count
the burden of	0.000	49
by and large	0.000	37
face of it	0.000	34
frame of mind	0.000	23
points of view	0.000	20
hair and a	0.000	17
the to infinitive	0.000	15
of alcohol and	0.000	8
a great many	0.083	44
glance up at	0.083	33
for and against	0.086	21
from of government	0.142	6

Example of Low Parsability N-grams

N-gram	R	Count
the burden of	0.000	49
by and large	0.000	37
face of it	0.000	34
frame of mind	0.000	23
points of view	0.000	20
hair and a	0.000	17
the to infinitive	0.000	15
of alcohol and	0.000	8
a great many	0.083	44
glance up at	0.083	33
for and against	0.086	21
from of government	0.142	6

[Zhang et al., 2006]

- Error mining based MWE detection
- New MWE entries created with automated lexical acquisition
- Grammar/Parser coverage improves significantly

? Validation steps are not thoroughly evaluated

? Grammar accuracy is not investigated

[Zhang et al., 2006]

- Error mining based MWE detection
- New MWE entries created with automated lexical acquisition
- Grammar/Parser coverage improves significantly

? Validation steps are not thoroughly evaluated

? Grammar accuracy is not investigated

Outline

Identification of MWEs

- Given a list of sequences of words to distinguish MWEs (e.g. *in the red*) from random sequences of words (e.g. *of alcohol and*)
- For statistical approaches there are two important questions
  - How reliable is the corpus used?
  - How precise is a statistical measure to distinguish the phenomena studied?

Resources

- 1039 trigrams from error mining system [van Noord, 2004]
- 4 corpora
  - BNC_f: fragment of the BNC used in the error-mining experiments
  - BNC: complete BNC (from the site http://pie.usna.edu/)
  - Google: Web using Google
  - Yahoo: Web using Yahoo

Corpus	Frequency of 1,039 trigrams
BNC_f	66,101
BNC	322,325
Google	224,479,065
Yahoo	6,081,786,313

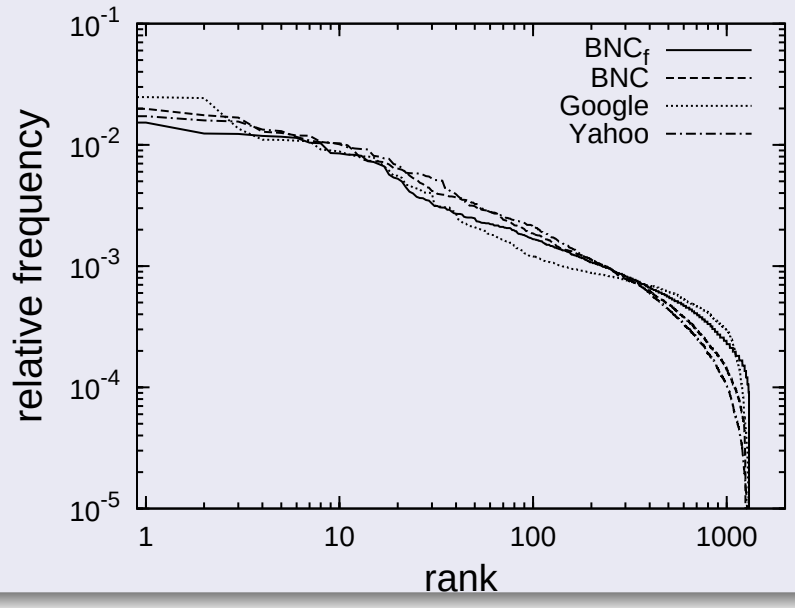
Comparing corpora

Hypothesis

The relative ordering in frequency for different n-grams is preserved across corpora, in the same domain
If not, different conclusions may be drawn from different corpora

Comparing corpora – first test

Relative Frequency Rank for the Trigrams



- The overall ranking distribution is very similar for these corpora, showing the expected Zipf like behaviour

Comparing corpora – second test

- Measuring Kendall's τ scores between corpora a significant correlation was found with $p < 0.000001$
- But what is the degree of correlation among them?
  - To estimate the correlation: the probability Q that any 2 trigrams chosen from two corpora have the same relative ordering in frequency

Comparing corpora – second test

	BNC	Google	Yahoo
BNC_f	0.81	0.73	0.78
BNC		0.73	0.77
Google			0.86

- The corpora are correlated, and can probably be used interchangeably for the statistical properties of the trigrams
- A higher correlation was observed between Yahoo and Google
  - It seems that as corpora sizes increase, so do the correlations between them

Comparing statistical measures

- Using a single corpus: BNC_f
- Comparing Mutual Information (MI), χ^2 and Permutation Entropy (PE) for MWE identification
- MI and χ^2 are typical measures of association that compare
  - the joint probability of occurrence of a certain group of events $p(abc)$
  - with a prediction derived from the *null* hypothesis of statistical independence between these events $p_\emptyset(abc) = p(a) \cdot p(b) \cdot p(c)$

MI and χ^2

$$\chi^2 = \sum_{a,b,c} \frac{[\, n(abc) - n_\emptyset(abc) \,]^2}{n_\emptyset(abc)}$$

$$\mathrm{MI} = \sum_{a,b,c} \frac{n(abc)}{N} \log_2 \left[\frac{n(abc)}{n_\emptyset(abc)} \right]$$

- a is the word w_1 (or $\neg w_1$), ...
- $n(a)$ is the number of unigrams a
- N is the number of words in the corpus
- $n(abc)$ is the number of trigrams abc in the corpus
- $n_\emptyset(abc) = n(a)n(b)n(c)/N^2$ is the predicted number from the *null* hypothesis

Permutation Entropy (PE)

- Permutation entropy, is a measure of order association

$$PE = -\sum_{(i,j,k)} p(w_i w_j w_k) \ln [\, p(w_i w_j w_k) \,]$$

$$p(w_1 w_2 w_3) = \frac{n(w_1 w_2 w_3)}{\sum_{(i,j,k)} n(w_i w_j w_k)}$$

- where the sum runs over all the permutations: (e.g. *by and large, large by and, and large by, and by large, large and by*, and *by large and*)

Permutation Entropy (PE)

- PE for MWE detection - Hypothesis: MWEs are more rigid to permutations; therefore they have smaller PEs
- the more independent the words are the closer PE is from its maximal value (ln 6, for trigrams)
- It does not rely on single word counts, which are less accurate in Web based corpora

Are they equivalent?

- Kendall's τ for assessing the correlation of the rankings for these measures and its significance
- Q is the probability of finding the same ordering in them

	MI×χ^2	MI×PE	χ^2×PE
Q	0.71	0.55	0.45

- The correlations found are statistically significant
- The measures order the trigrams differently
  - 70% chance of getting the same order from MI and χ^2
  - they are very different from the PE

Are they useful for MWE detection?

- To check that we compare the measures' distributions for MWEs and non-MWEs
- Gold standard = set of 382 MWE candidates annotated by a native speaker
  - 90 MWEs
  - 292 non-MWEs

Are they useful?

Kolmogorov-Smirnof Test

- D value (D$\in$[0,1]): large values indicate large differences between distributions
- p: significance probability associated to D

	MI_{BNC_f}	$\chi^2_{BNC_f}$	PE_{Yahoo}	PE_{Google}
D	0.27	0.13	0.27	0.24
p<	0.0001	0.154	0.0001	0.0005

- MI or PE seem to differentiate between MWEs and non-MWEs

Are they useful?
Normalised histograms for MWEs and non-MWEs

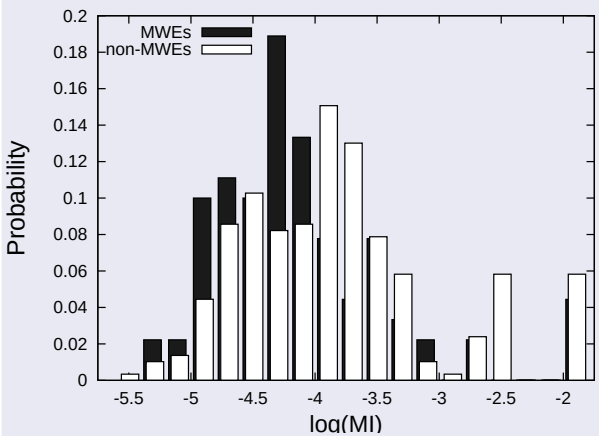
- The ideal scenario: non overlapping distributions for MWEs and non-MWEs
  - A simple threshold operation would be enough to distinguish between them

Are they useful?
Normalised histograms for MWEs and non-MWEs

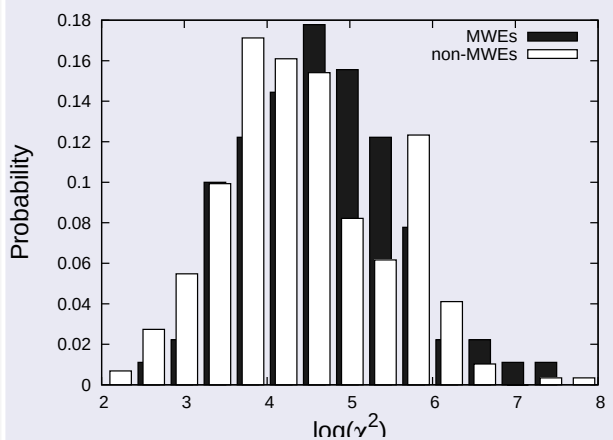
- The ideal scenario: non overlapping distributions for MWEs and non-MWEs
  - A simple threshold operation would be enough to distinguish between them

Are they useful?
Normalised histograms for MWEs and non-MWEs

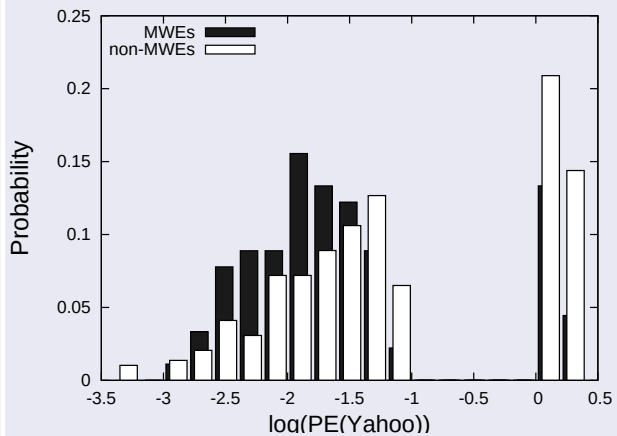
MI (BNC_f)



χ^2 (BNC_f)



PE (Yahoo)



- As some types of MWEs may have stronger constraints on word order, more visible effects will probably be seen if we look at application of measures for individual types of MWEs [Evert and Krenn, 2005]

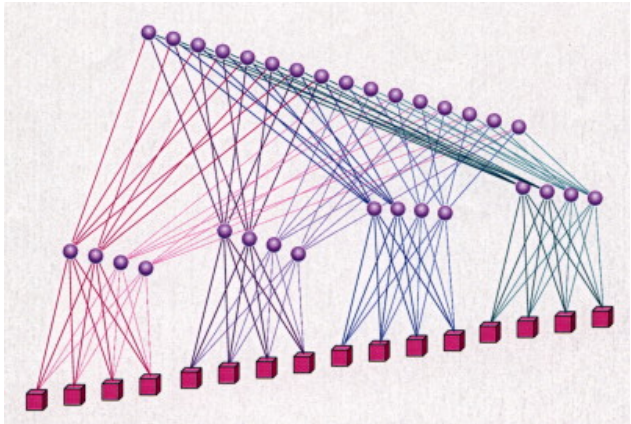
Outline

English Resource Grammar [Flickinger, 2000]

- A large scale broad coverage precision HPSG grammar
- Lexicon coverage is a major problem
- MWEs comprise a large portion of the missing lexical entries

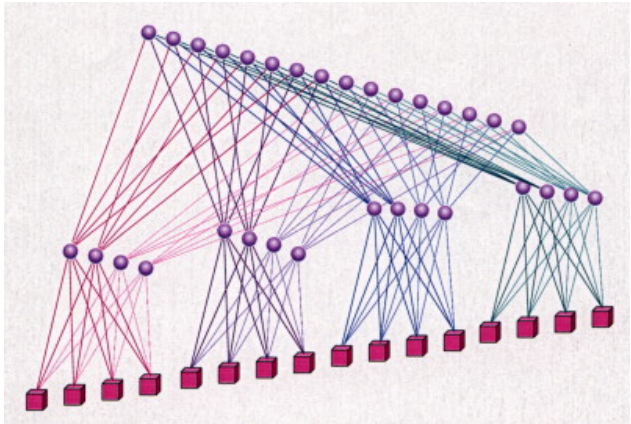
Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



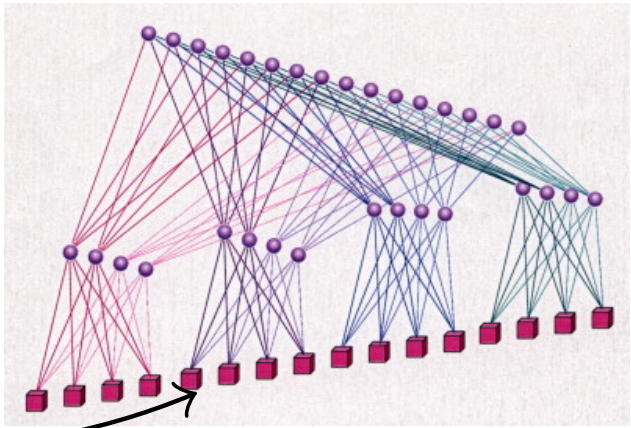
Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



Lexical hierarchy and atomic lexical types

- The lexical information is encoded in atomic lexical types
- A lexicon is a $n : n$ mapping between lexemes and atomic lexical type



Maximum Entropy Model-based Lexical Type Predictor

- A statistical classifier that predicts for each occurrence of an unknown word or a missing lexical entry
- Input: features from the context
- Output: atomic lexical types

$$p(t,c) = \frac{exp(\sum_i \theta_i f_i(t,c))}{\sum_{t' \in T} exp(\sum_i \theta_i f_i(t',c))}$$

"Words-with-spaces" vs. compositional approaches

Words-with-spaces approach [Zhang et al., 2006]

- Assign lexical types for the entire MWE
- Grammar coverage significantly improves
- Grammar accuracy decreases

Compositional approach

- Assign new lexical entries for the head word to treat the MWE as compositional
- Hopefully the grammar coverage improves without drop in accuracy

"Words-with-spaces" vs. compositional approaches

Words-with-spaces approach [Zhang et al., 2006]

- Assign lexical types for the entire MWE
- Grammar coverage significantly improves
- Grammar accuracy decreases

Compositional approach

- Assign new lexical entries for the head word to treat the MWE as compositional
- Hopefully the grammar coverage improves without drop in accuracy

Experiment

- Rank all the MWE candidates according to the three statistical measures: MI, χ^2, PE, and select the top 30 MWE with highest average ranking
- Extract sub-corpus from BNC_f which contains at least one of the MWE for evaluation (674 sentences)
- Use heuristics to extract head words (20 head words)
- Run lexical acquisition for head words on the sub-corpus (21 new entries)

Grammar Coverage

	item #	parsed #	avg. analysis #	coverage %
ERG	674	48	335.08	7.1%
ERG + MWE	674	153	285.01	22.7%

- The coverage improvement is largely compatible with the results of “words-with-spaces” approach reported in [Zhang et al., 2006] (about 15%)
- Great reduction in lexical entries added

Grammar Accuracy

- 153 parsed sentences are analyzed by hand
- 124 (81.0%) of them receive at least one correct/acceptable analysis (comparable to the accuracy reported by [Baldwin et al., 2004])
- Parse selection model finds best analysis in top-5 for 66% of the cases, and top-10 for 75%

Summary

- Different corpora are compared for the purpose of MWE validation
- Different statistical measures are compared for identifying MWEs
- Grammar performance for robust deep parsing is evaluated for automated MWE acquisition using compositional approach

Outlook

- Hand-crafted precision grammars usually face coverage/robustness challenges when applied to unseen data with unknown words/MWEs, unknown constructions, etc., all over the place
- [Baldwin et al., 2004] reported parsing coverage of **18**% on unseen BNC data parsed with the ERG, with the majority of parsing failures related to missing lexical entries
- The Lexical Type Prediction model I have presented above is used to handle unknown words (simplex and MWE) on-the-fly
- With the use of this model the ERG achieves around **84**% parsing coverage on unseen WSJ data

Outlook

Other "Deep" Parsing Systems

- LFG
  - XLE 79.6% F-Score [Kaplan et al., 2004]
- CCG
  - C&C 81.86% F-Score [Clark and Curran, 2007]
- HPSG
  - Enju 82.64% F-Score [Sagae et al., 2008]
- The aforementioned systems are evaluated on 700 sentences selected from WSJ data (PARC 700), using Grammatical Relations (GR)

For Further Reading I

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., and Oepen, S. (2004).

Road-testing the English Resource Grammar over the British National Corpus.
In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Clark, S. and Curran, J. (2007).

Formalism-Independent Parser Evaluation with CCG and DepBank.
In *Proceedings of ACL2007*.

Evert, S. and Krenn, B. (2005).

Using small random samples for the manual evaluation of statistical association measures.
Computer Speech and Language, 19(4):450–466.

Flickinger, D. (2000).

On building a more efficient grammar by exploiting types.
Natural Language Engineering, 6(1):15–28.

Jackendoff, R. (1997).

Twistin' the night away.
Language, 73:534–59.

Kaplan, R., Riezler, S., King, T. H., Maxwell, J., and Vasserman, A. (2004).

Speec and accuracy in shallow and deep stochastic processing.
In *Proceedings of HLT-NAACL'04*.

For Further Reading II

Lucchesi, C. and Kowaltowski, T. (1993).

Applications of finite automata representing large vocabularies.
Software Practice and Experience, 23(1):15–30.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).

Multiword expressions: A pain in the neck for NLP.
In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Sagae, K., Miyao, Y., Matsuzaki, T., and Tsujii, J. (2008).

Challenges in Mapping of Syntactic Representations for Framework-Independent Parser Evaluation.
In *Proceedings of Workshop on Automated Syntatic Annotations for Interoperable Language Resources at the First International Conference on Global Interoperability for Language Resources (ICGL'08)*, Hong Kong.

van Noord, G. (2004).

Error mining for wide-coverage grammar engineering.
In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 446–453, Barcelona, Spain.

Villavicencio, A. (2005).

The availability of verb-particle constructions in lexical resources: How much is enough?
Journal of Computer Speech and Language Processing, 19.

For Further Reading III

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006).

Automated multiword expression prediction for grammar engineering.

In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.