Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources

Yulia Tsvetkov and Shuly Wintner

Department of Computer Science
University of Haifa
Haifa, Israel

January 24, 2011

Outline

Multiword Expressions

Characterization

Mutli-word expressions (MWEs) are lexical words consisting of more than a single orthographic word

- Morphologically, their behavior is often idiosyncratic
- Syntactically, they function as words or as phrases
- Semantically, their meaning is usually non-compositional (i.e., cannot be established from the meanings of their components)

MWEs blur the boundaries between the lexicon and the grammar

Multiword Expressions

Scale

- MWEs constitute a major part of any language, and the magnitude of this phenomenon is far greater than has traditionally been realized within linguistics
- Jackendoff (1997, page 156) estimates that the number of MWEs in a speakers' lexicon (in English) is of the same order of magnitude as the number of single words
- In WordNet 1.7 (Fellbaum, 1998), 41% of the entries are multiwords
- Erman and Warren (2000) revealed that over 55% of the tokens in the texts they studied were instances of what they call *prefabs*

Multiword Expressions

Significance

- Identification of MWEs is an important task for a variety of natural language processing (NLP) applications (Villavicencio et al., 2005):
  - Information retrieval (Doucet and Ahonen-Myka, 2004)
  - Building ontologies (Venkatsubramanyan and Perez-Carballo, 2004)
  - Text alignment (Venkatapathy and Joshi, 2006)
  - Machine translation (Baldwin and Tanaka, 2004; Uchiyama et al., 2005)
- Expressions with idiosyncratic features that cannot be predicted on the basis of their component words must be included in language descriptions (such as lexicons) in order to account for actual usage

Research Goal

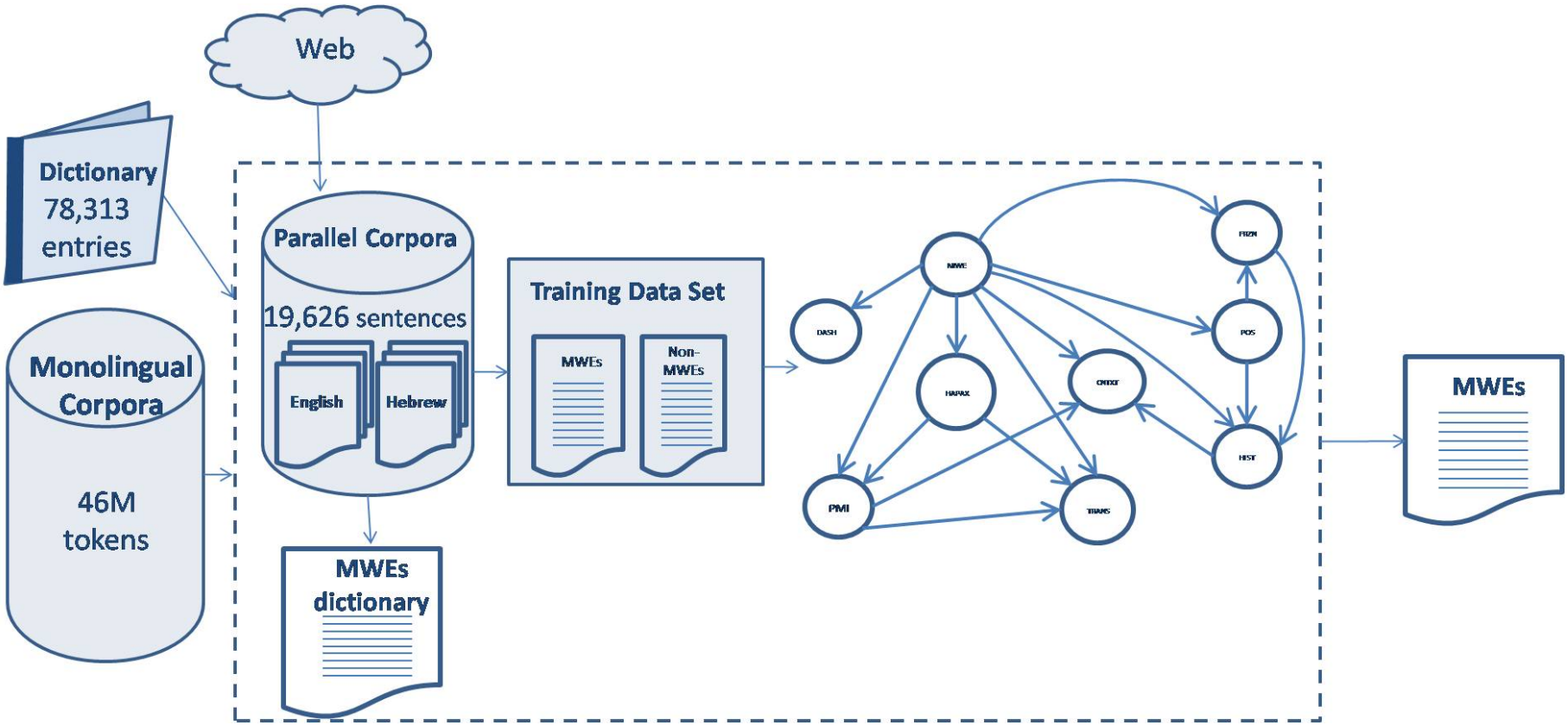
Develop techniques to extract MWEs from corpora, and populate the lexicon with MWEs acquired automatically

- MWEs of various types and syntactic categories
- Language-independent
- Appropriate for medium-density languages with few language resources

Related work

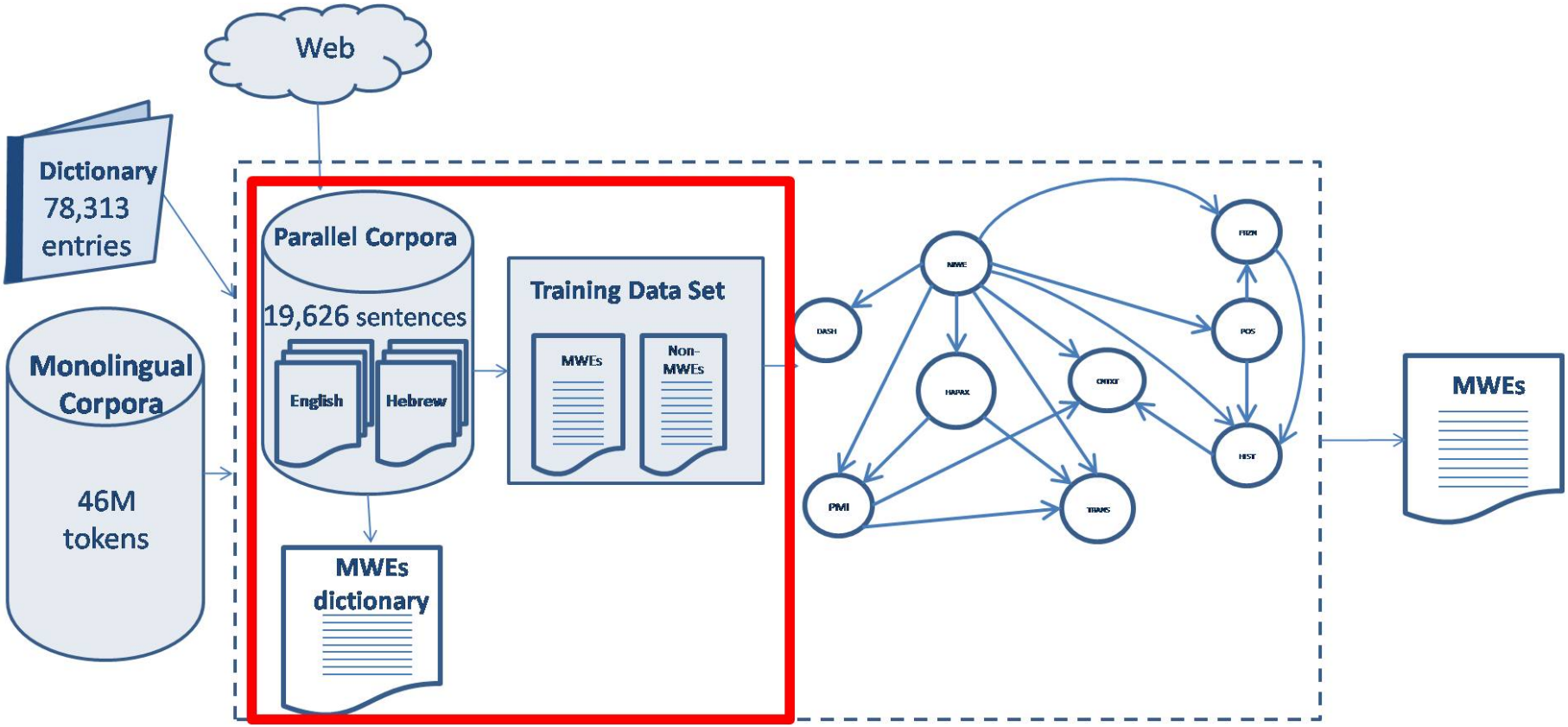
- Early approaches concentrated on the collocational behavior of MWEs (Church and Hanks, 1989)
- Combining different collocation measures significantly improves performance (Pecina, 2008)
- Adding linguistic information to collocation measures can improve identification accuracy:
  - Adding POS patterns helps to identify English Verb-Particle Constructions and German Adjective-Noun pairs (Ramisch et al., 2008)
  - Van de Cruys and Villada Moirón (2007) use lexical fixedness to extract Dutch Verb-Noun idiomatic combinations; Bannard (2007) uses syntactic fixedness to identify them in English; Fazly and Stevenson (2006) use both
- Various properties of Hebrew MWEs are described by Al-Haj (2010); Al-Haj and Wintner (2010) use them in order to construct an SVM-based classifier that can distinguish between MWE and non-MWE Noun-Noun constructions in Hebrew.

Architecture



- Monolingual corpora: MILA corpus and tools: (Itai and Wintner, 2008) (Bar-Haim et al., 2005)
- Bilingual dictionary (Itai and Wintner, 2008; Kirschenbaum and Wintner, 2010)
- Small parallel corpus (Tsvetkov and Wintner, 2010b)

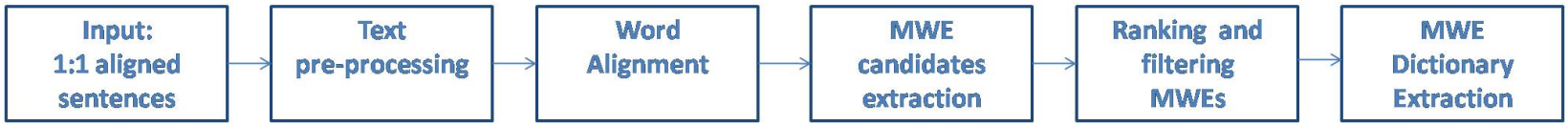
Acquisition of MWEs from parallel corpora



- Tsvetkov and Wintner (2010a)

Acquisition of MWEs from parallel corpora

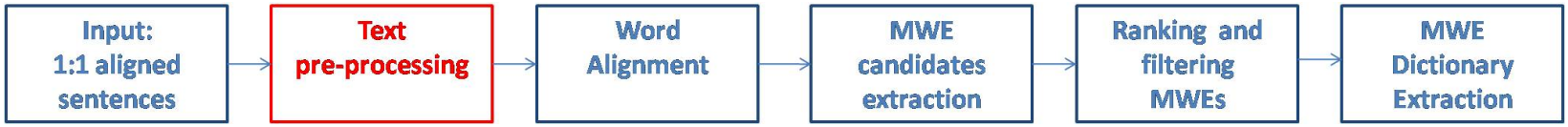
Motivation



- MWEs are aligned across languages in a way that differs from compositional expressions
  - *Mis*-alignments are candidate MWEs
- Parallel corpus is particularly small, and we cannot fully rely on the quality of word alignments
  - Candidate MWEs are ranked by statistics estimated from a large *monolingual* corpus

Acquisition of MWEs from parallel corpora

Preprocessing



We try to remove some language-specific differences automatically:

- Tokenization, morphological analysis and disambiguation
- The surface form of each word is reduced to its base form, and bound morphemes are split to generate stand-alone "words"
- Tokenization and lemmatization of the English side, using the NLTK package (Bird et al., 2009)
- Remove frequent function words that do not have direct counterparts in the other language: in English, the articles *a*, *an* and *the*, the infinitival *to* and the copulas *am*, *is* and *are*; in Hebrew, the accusative marker *at* and the definite article *h*

Acquisition of MWEs from parallel corpora

Example - Preprocessing

Example

A Hebrew sentence from our corpus:

wamrti	*lh*	*lhzhr*	*mbn*	*adm*	*kzh*
and-I-told	to-her	to-be-careful	from-child	man	like-this

"and I told her to keep away from the person"

After preprocessing:

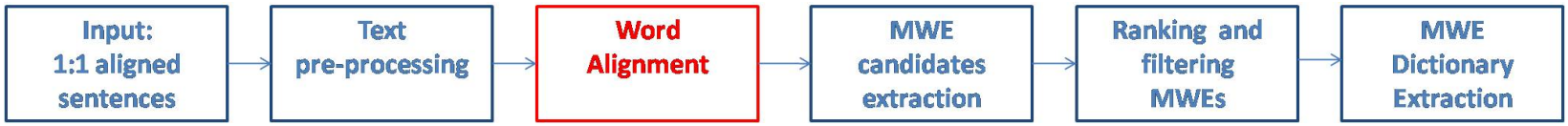
w	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn*	*adm*	*k*	*zh*
and	I	tell	to-her	to-be-careful	from	child	man	like	this

The English sentence is represented as:

and i tell her keep away from person

Acquisition of MWEs from parallel corpora

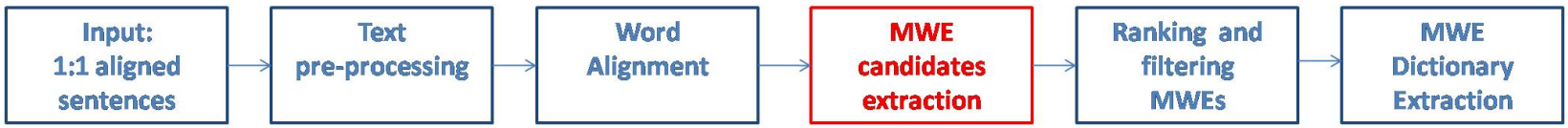
Word alignment



- Use Giza++ (Och and Ney, 2003) to word-align the text
- Look up all $1:1$ alignments in the bilingual dictionary

Acquisition of MWEs from parallel corpora

MWE candidates extraction



- If the pair exists in the dictionary, we remove it from the sentence and replace it with a special symbol, '*'. Such word pairs are not parts of MWEs
- If the pair is not in the dictionary, but its alignment score as produced by Giza++ is high and it is sufficiently frequent, we add the pair to the dictionary but also retain it in the sentence. Such pairs are still candidates for being (parts of) MWEs

Acquisition of MWEs from parallel corpora

Example - MWE candidates extraction

Example

Giza++ alignments:

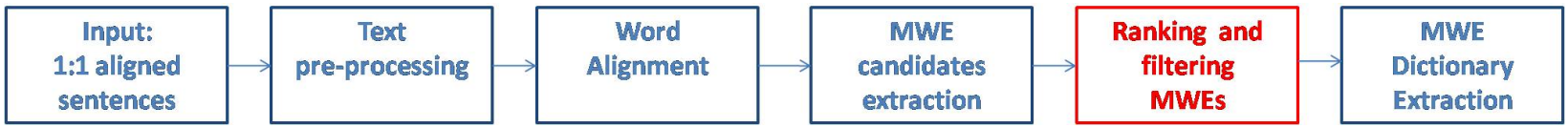
w	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn adm*	*k*	*zh*
and	I	told	her	keep away	from	person	{}	{}

Once 1:1 alignments are replaced by '*':

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

Ranking and filtering MWEs



- Extract all word bi-grams from the candidates. Each bi-gram is associated with its PMI-based score, computed from the *monolingual corpus*.
- A word sequence of any length is considered MWE if all its adjacent bi-grams score above the threshold.
- Restore the original forms of the Hebrew words in the candidates; restore the function words.

Acquisition of MWEs from parallel corpora

Example - Ranking and filtering MWEs

Example

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

Example - Ranking and filtering MWEs

Example

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

Example - Ranking and filtering MWEs

Example

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

Example - Ranking and filtering MWEs

Example

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

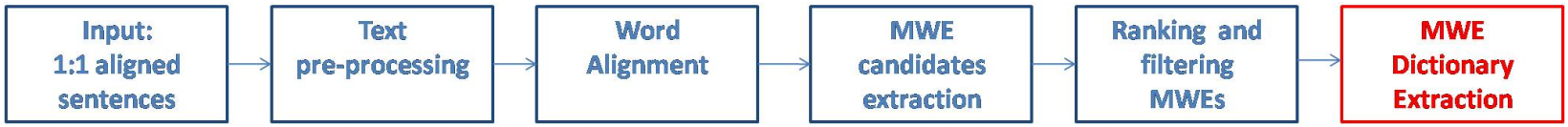
Example - Ranking and filtering MWEs

Example

*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
*	*	*	*	keep away	*	person		

Acquisition of MWEs from parallel corpora

Dictionary extraction



- For each MWE in the source-language sentence, we consider as translation all the words in the target-language sentence (in their original order) that are aligned to the word constituents of the MWE, as long as they form a contiguous string.
- The result is a bilingual dictionary containing 2,955 MWE translation pairs, and also 355 translation pairs produced by taking high-quality 1:1 word alignments.

Acquisition of MWEs from parallel corpora

Results: 15 top-ranking extracted MWEs

Hebrew	Gloss	Type
xbr hknst	Member of Parliament	NNC
tl abib	Tel Aviv	GT
gwš qTip	Gush Katif	NNC-GT
awpir pins	Ophir Pines	PN
hc't xwq	Legislation	NNC
axmd Tibi	Ahmad Tibi	PN
zhwh glawn	Zehava Galon	PN
raš hmmšlh	Prime Minister	NNC
abšlwm wiln	Avshalom Vilan	PN
br awn	Bar On	PN
mair šTrit	Meir Shitrit	PN
limwr libnt	Limor Livnat	PN
hiw'c hmšpTi	Attorney General	N-ADJ
twdh rbh	thanks a lot	N-ADJ
rcw't 'zh	Gaza Strip	NNC-GT

Acquisition of MWEs from parallel corpora

Results

Of the top 100 candidates, 99 are clearly MWEs, including:

- *mzg awir* (*temper-of air*) "weather"
- *kmw kn* (*like thus*) "furthermore"
- *bit spr* (*house-of book*) "school"
- *šdh t'wph* (*field-of flying*) "airport"
- *tšwmt lb* (*input-of heart*) "attention"
- *ai apšr* (*not possible*) "impossible"
- *b'l ph* (*in-on mouth*) "orally"

Longer MWEs include:

- *ba lidi biTwi* (*came to-the-hands-of expression*) "was expressed"
- *xzr 'l 'cmw* (*returned on itself*) "recurred"
- *ixd 'm zat* (*together with it*) "in addition"
- *h'crt hkllit šl haw"m* (*the general assembly of the UN*) "the UN general assembly"

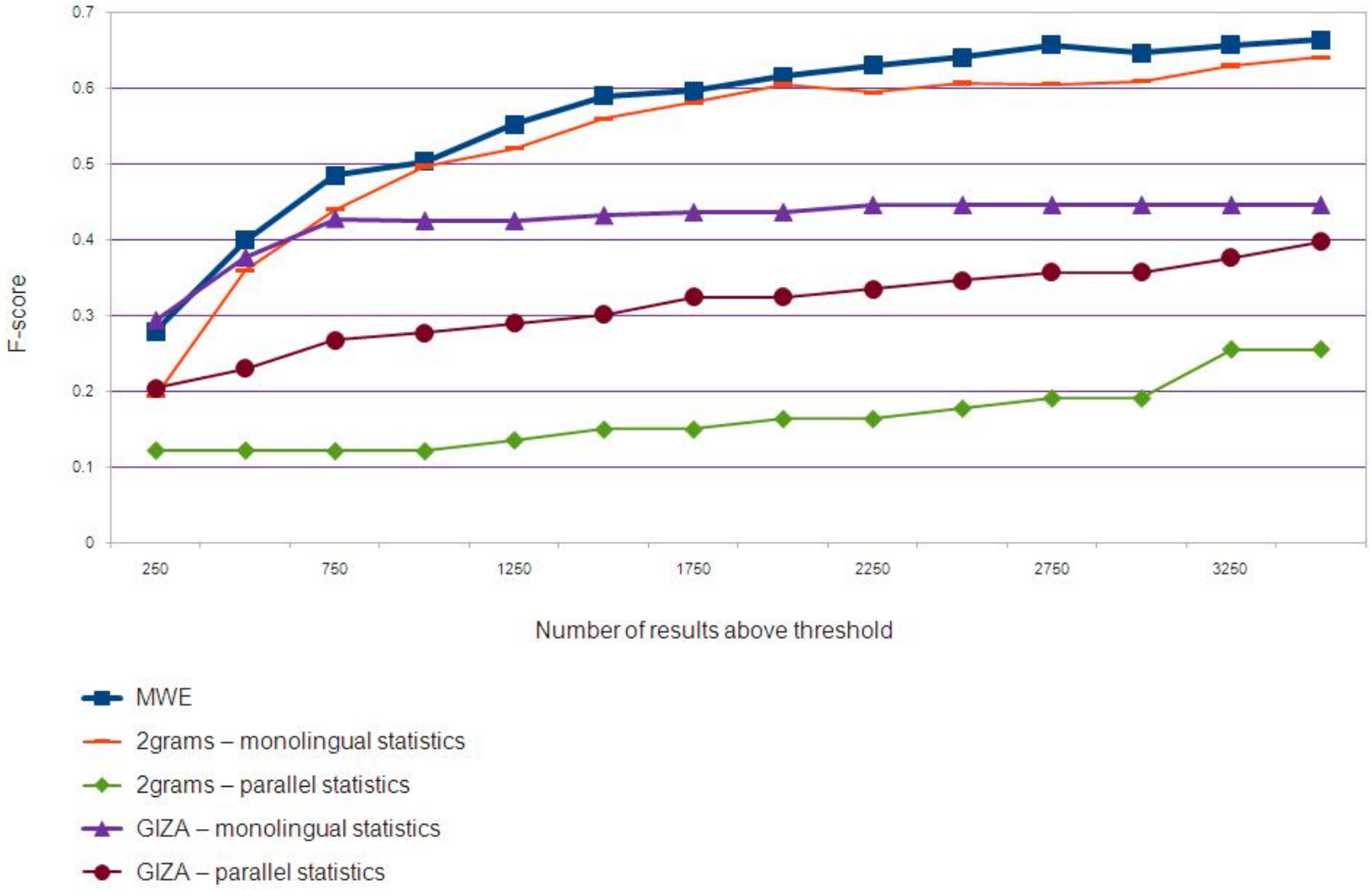
Acquisition of MWEs from parallel corpora

Evaluation

- Of the noun-noun construction set (Al-Haj and Wintner, 2010), 121 positive and 91 negative examples are included in our corpus
- We compare our results to three baselines:
  - using only PMI to rank the bi-grams in the parallel corpus
  - using PMI computed from the monolingual corpus to rank the bi-grams in the parallel corpus
  - using Giza++ 1:n alignments, ranked by their PMI (with bi-gram statistics computed once from parallel and once from monolingual corpora)

Acquisition of MWEs from parallel corpora

Evaluation



Acquisition of MWEs from parallel corpora

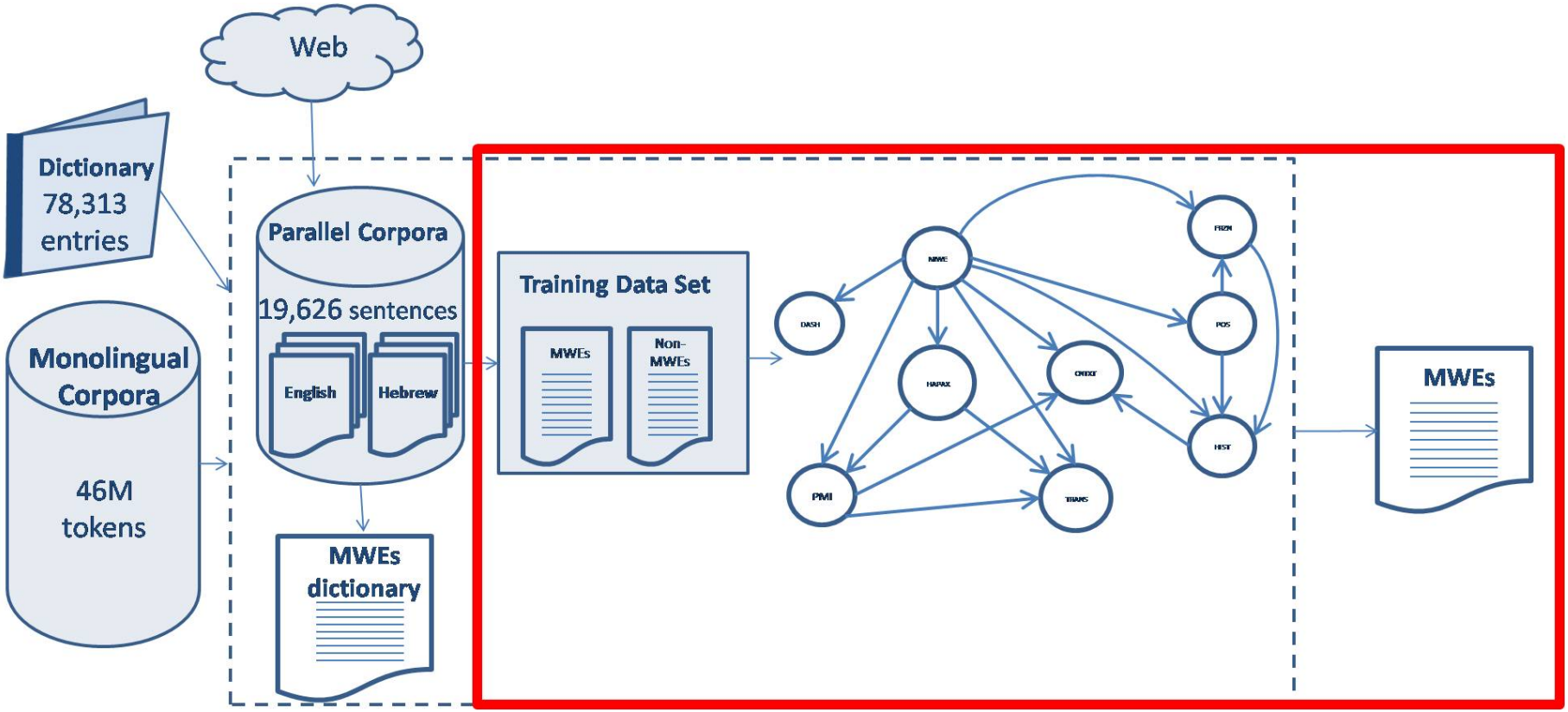
Evaluation

- We used the extracted MWEs, *along with their translations*, to augment the lexicon of a Hebrew to English machine translation system (Lavie et al., 2004)
- We added 2,955 MWE translation pairs, plus 355 translation pairs reflecting high-quality $1:1$ word alignments
- The results reflect a small but statistically significant improvement in the performance of the system:

Dictionary	BLEU	Meteor
Original	13.69	33.38
Augmented	13.79	33.99

Extraction of MWEs from Monolingual Corpora

Algorithm



Bayesian Network

Motivation

Bayesian Networks provide an optimal architecture for expressing various pieces of knowledge aimed at MWE identification:

- BN models tend to be more stable and handle missing or noisy data better.
- Causal relationships between features facilitate better understanding of the problem domain.
- Encode not only statistical data, but also prior domain knowledge and human intuitions, in the form of interdependencies among features.

Linguistically-motivated Features

We define a set of linguistically-motivated features, in which we attempt to capture **idiosyncratic behavior** of MWEs along following dimensions (Al-Haj and Wintner, 2010):

- Morphological
- Semantic
- Syntactic
- Statistical

Morphological idiosyncrasy

Orthographic variation

Description:

- Sometimes, MWEs are written with dashes instead of inter-token spaces.

Computable equivalent:

- We define a binary feature, DASH, whose value is 1 iff the dash character appears in some surface form of the candidate MWE.

Example

xd-cddi (*one sided*) "unilateral".

Morphological idiosyncrasy

Hapax legomena

Description:

- MWEs sometimes include constituents that have no usage outside the particular expression, and are hence not included in lexicons.

Computable equivalent:

- We define a feature, HAPAX, whose value is a binary vector with 1 in the *i*-th place iff the *i*-th word of the candidate is not in the lexicon, and does not occur in other bi-grams at the same location.

Example

hwqws pwqws "hocus-pocus".

Morphological idiosyncrasy

Frozen form

Description:

- MWE constituents sometimes occur in one fixed, frozen form.

Computable equivalent:

- We define a feature, FROZEN, whose value is a binary vector with 1 in the *i*-th place iff the *i*-th word of the candidate never inflects in the context of this expression.

Example

bit xwlim (*house-of sick-people*) "hospital"

- the noun *xwlim* must be in the plural in this MWE.

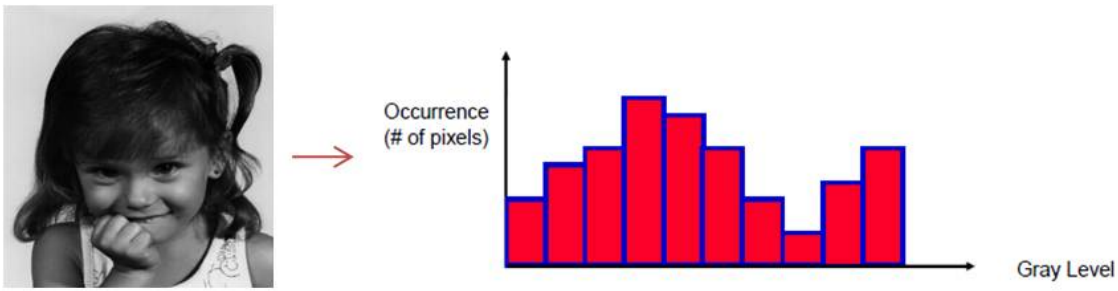
Morphological idiosyncrasy

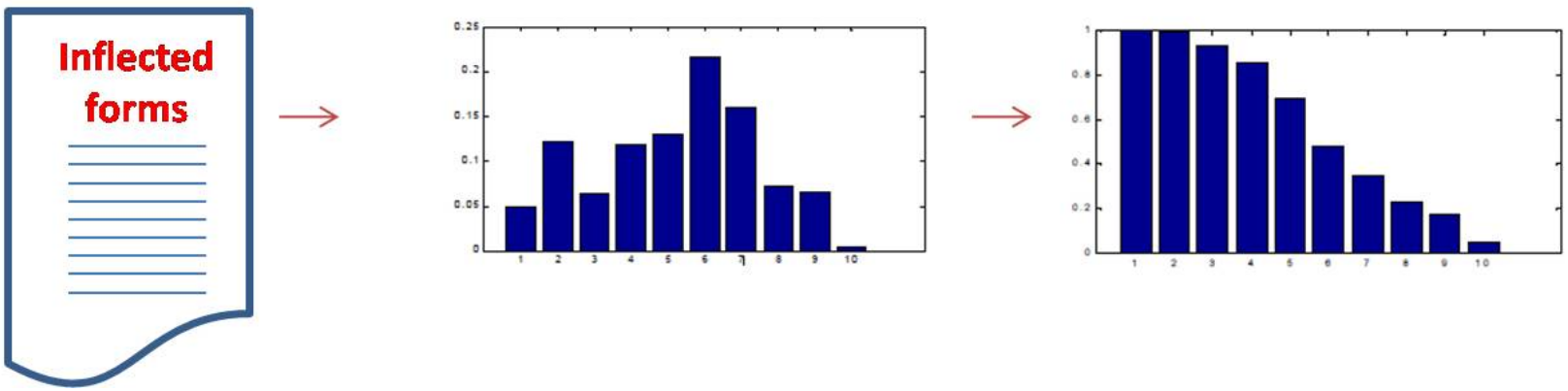
Partial morphological inflection

Description:

- In some cases, MWE constituents undergo a (strict but non-empty) subset of the full inflections that they would undergo in isolation.

Computable equivalent:

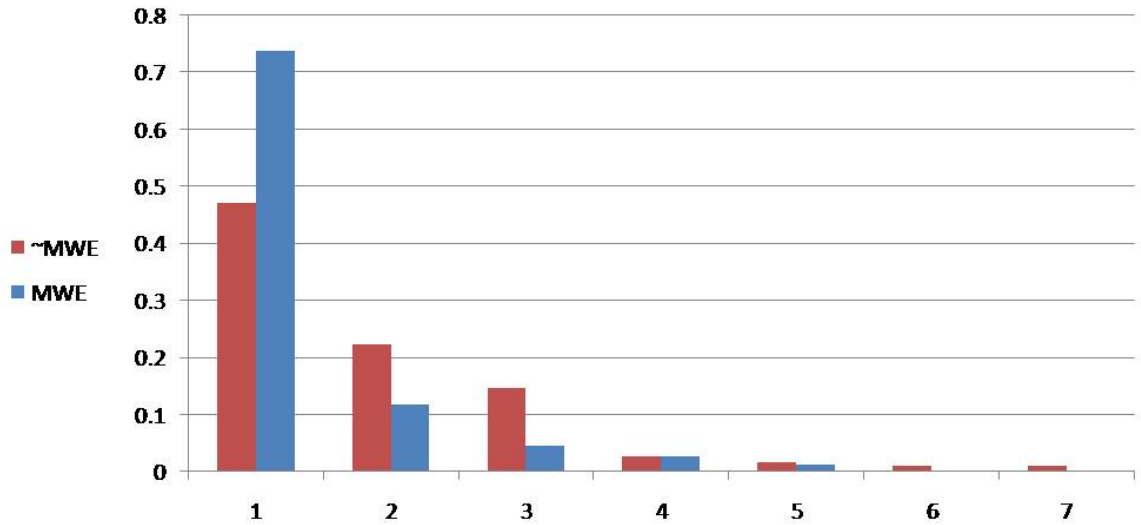




Morphological idiosyncrasy

Partial morphological inflection

The average histogram for positive and negative examples:



- $H(i)$ specifies the # of occurrences of inflected form i
- Let N be the number of inflected forms $N = \sum_i H(i)$
- Normalized histogram: $P(i) = H(i)/N$
- Average histograms for positive and negative examples

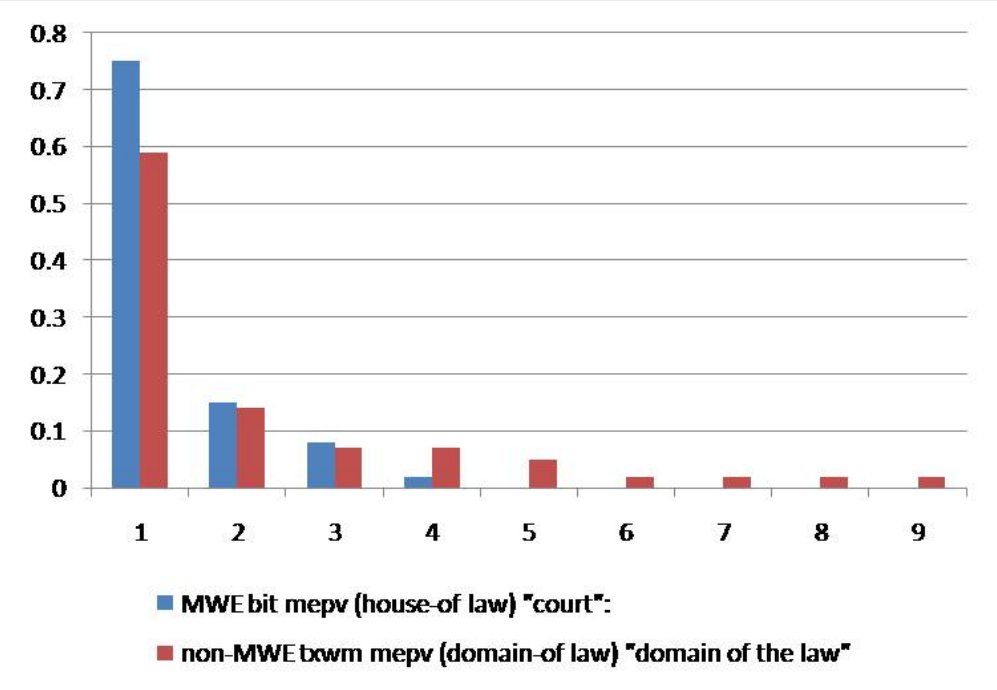
Morphological idiosyncrasy

Partial morphological inflection

Distance measure:

- $Dist(phrase1, phrase2) = \sum_i \left| P_{phrase1}(i) - P_{phrase2}(i) \right|$
- We define as feature, HIST, the L_1 (Manhattan) distance between the histogram of the candidate and the closest average histogram.

Example



Semantic idiosyncrasy

Translational equivalents

Description:

- Since MWEs are often idiomatic, they tend to be translated in a non-literal way, sometimes to a single word.

Computable equivalent:

- We use a dictionary to generate word-by-word translations of candidate MWEs to English, and check the number of occurrences of the English literal translation in a large English corpus.
- We define a binary feature, TRANS, whose value is 1 iff some literal translation of the candidate occurs in the corpus.

Example

htxtn ym (*marry with*) "marry"

- literally translated as *with marry*, *marry with*, *together marry* and *marry together*, none of which occurs in the corpus.

Semantic idiosyncrasy

Context

Description:

- We hypothesize that MWEs tend to constrain their syntactic context more strongly than compositional expressions.

Computable equivalent:

- We compute a histogram of the frequencies of words *following* each candidate MWE.
- We trim the tail of the histogram by removing low-frequency words (the expectation is that non-MWEs would have a much longer tail).
- Off-line, we compute the same histograms for positive and negative examples and average them.
- The value of CONTEXT is 1 iff the histogram of the candidate is closer (in terms of L_1 distance) to the positive average.

Syntactic diversity

Description:

- MWEs can belong to various part of speech categories.

Computable equivalent:

- We define as feature, POS, the category of the candidate, with values obtained by selecting frequent tuples of POS tags.

Example

Noun-Noun, PropN-PropN, Noun-Adj, etc.

Collocation

Description:

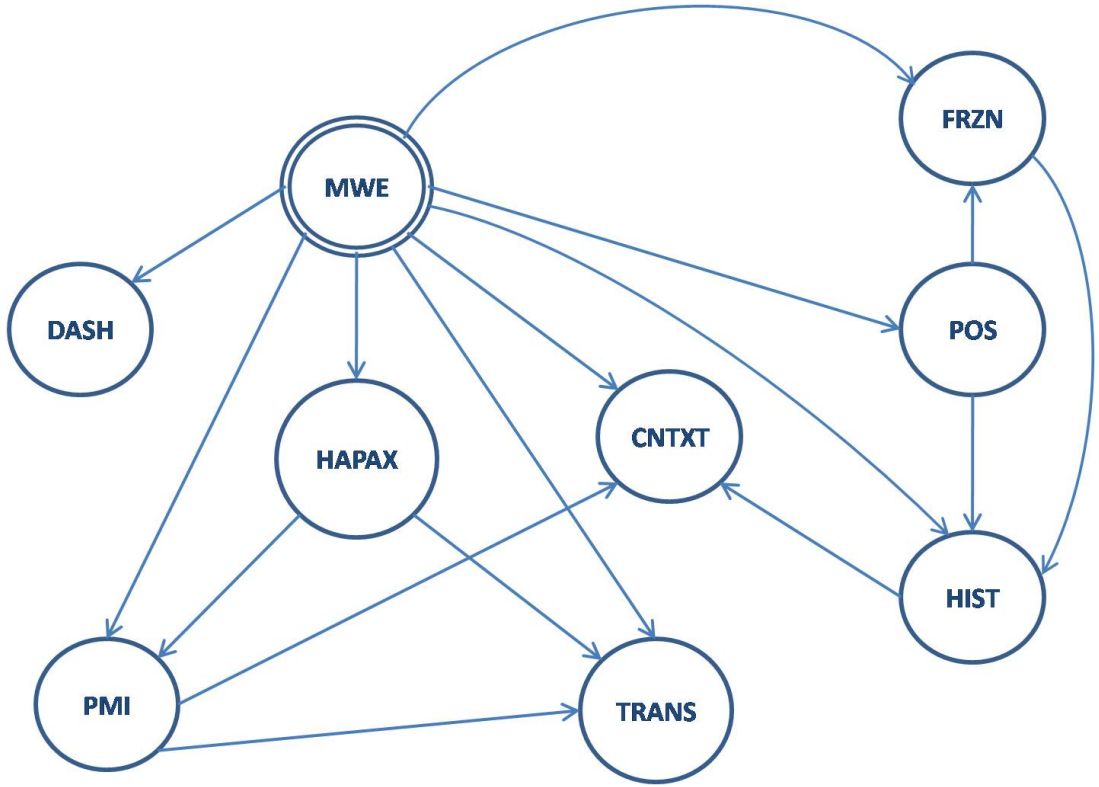
- As a baseline, collocation-based measure, we use PMI.

Computable equivalent:

- We define a binary feature, PMI, with values (*low* and *high*) reflecting the threshold that maximizes the accuracy of MWE classification in Tsvetkov and Wintner (2010a).

Bayesian Network for MWE identification

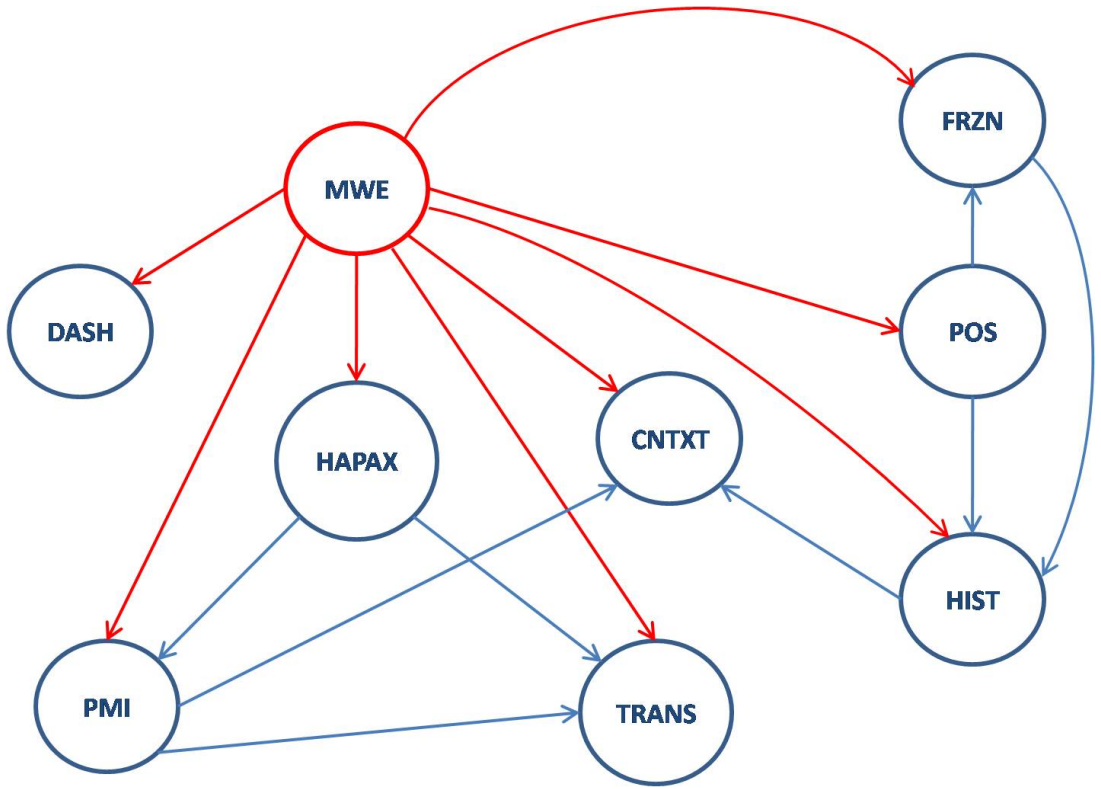
Feature Interdependencies



We use a particular type of BN, known as *causal* networks, in which directed edges lead to a variable from each of its direct *causes*. This facilitates the expression of domain knowledge (and intuitions, beliefs, etc.) as structural properties of the network.

Bayesian Network for MWE identification

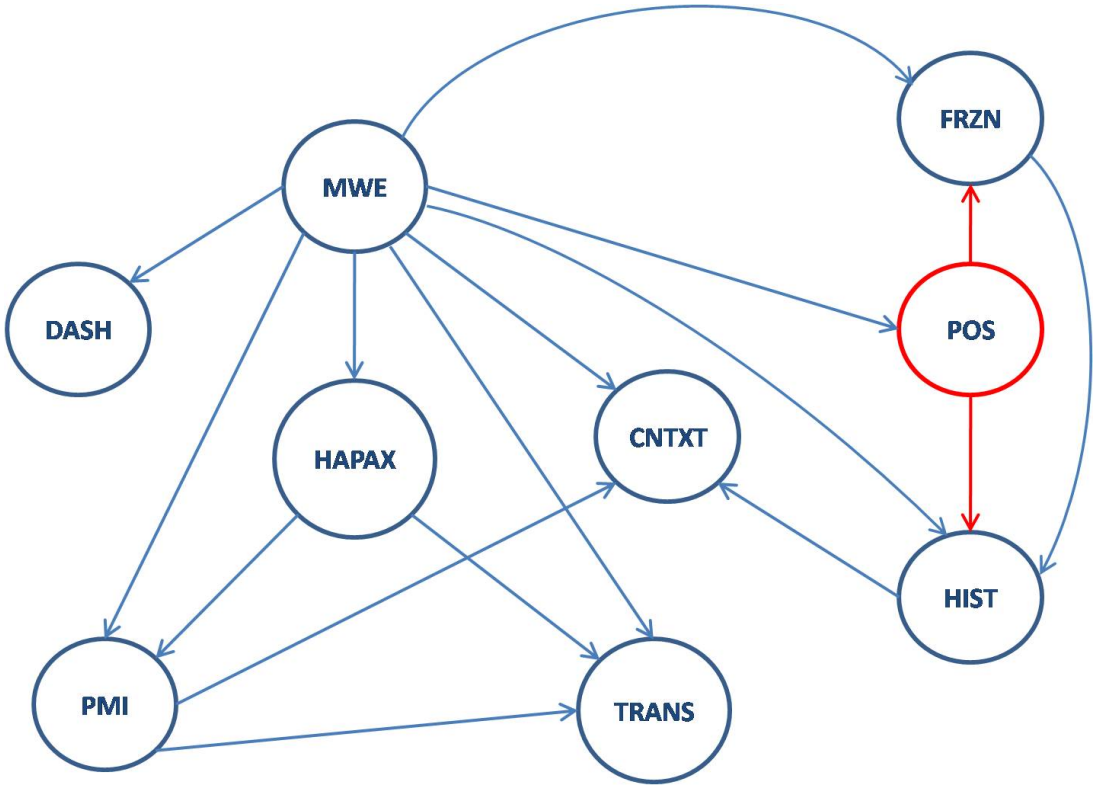
Feature Interdependencies



All nodes depend on MWE, as all are affected by whether or not the candidate is a MWE.

Bayesian Network for MWE identification

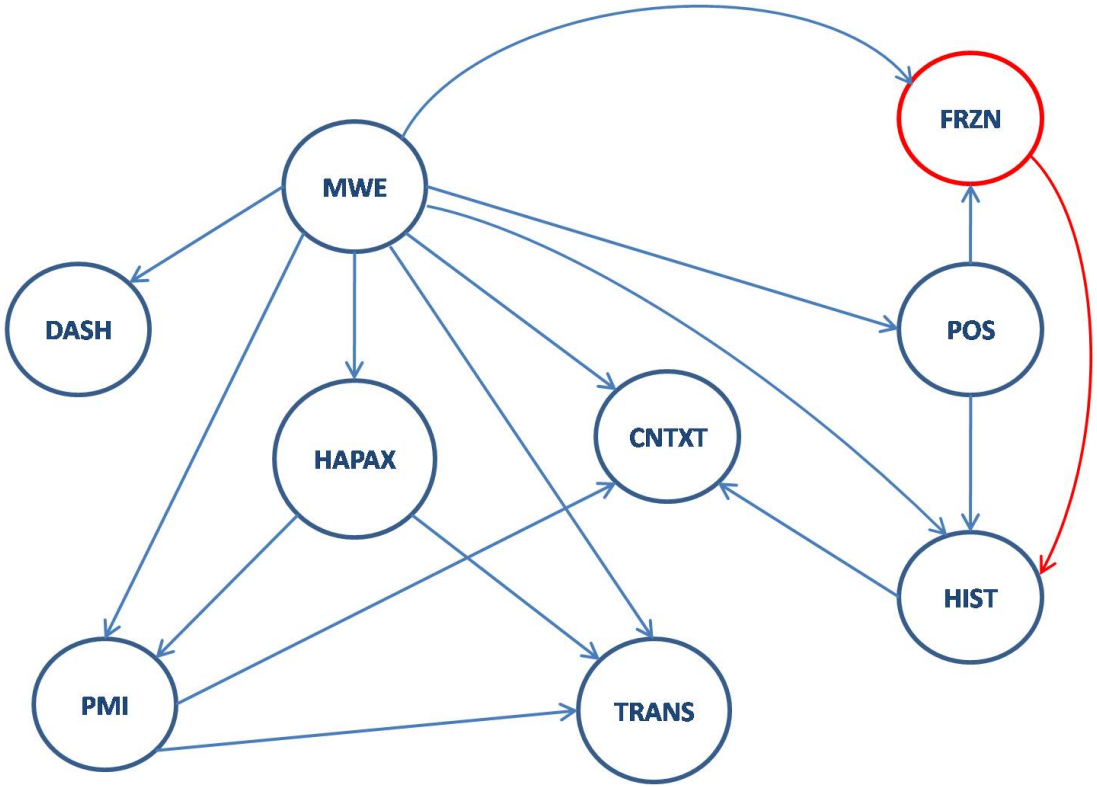
Feature Interdependencies



The POS of an expression influences its morphological inflection, hence the edges from POS to HIST and to FROZEN.

Bayesian Network for MWE identification

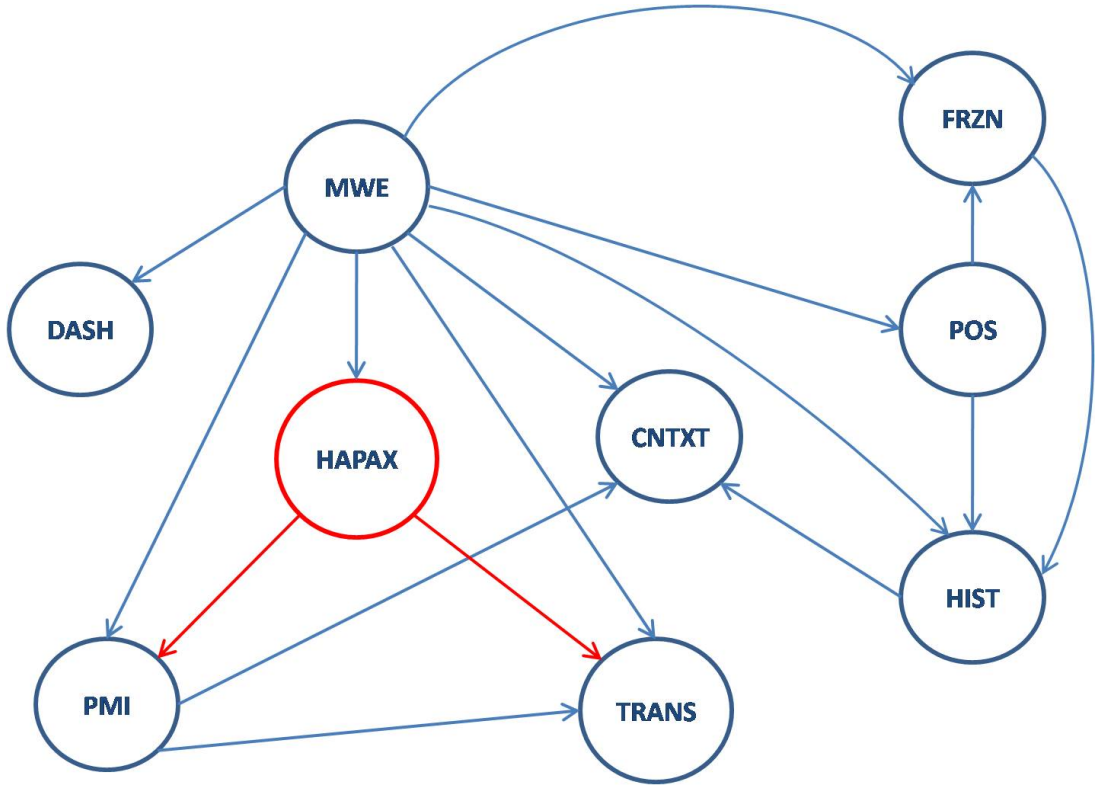
Feature Interdependencies



Clearly, FROZEN affects the distributional information on morphological behavior of MWE.

Bayesian Network for MWE identification

Feature Interdependencies



Hapaxes clearly affect all statistical metrics, hence the edge from HAPAX to PMI, and also the existence of literal translation, since if a word is not in the lexicon, it does not have a translation, hence the edge from HAPAX to TRANS.

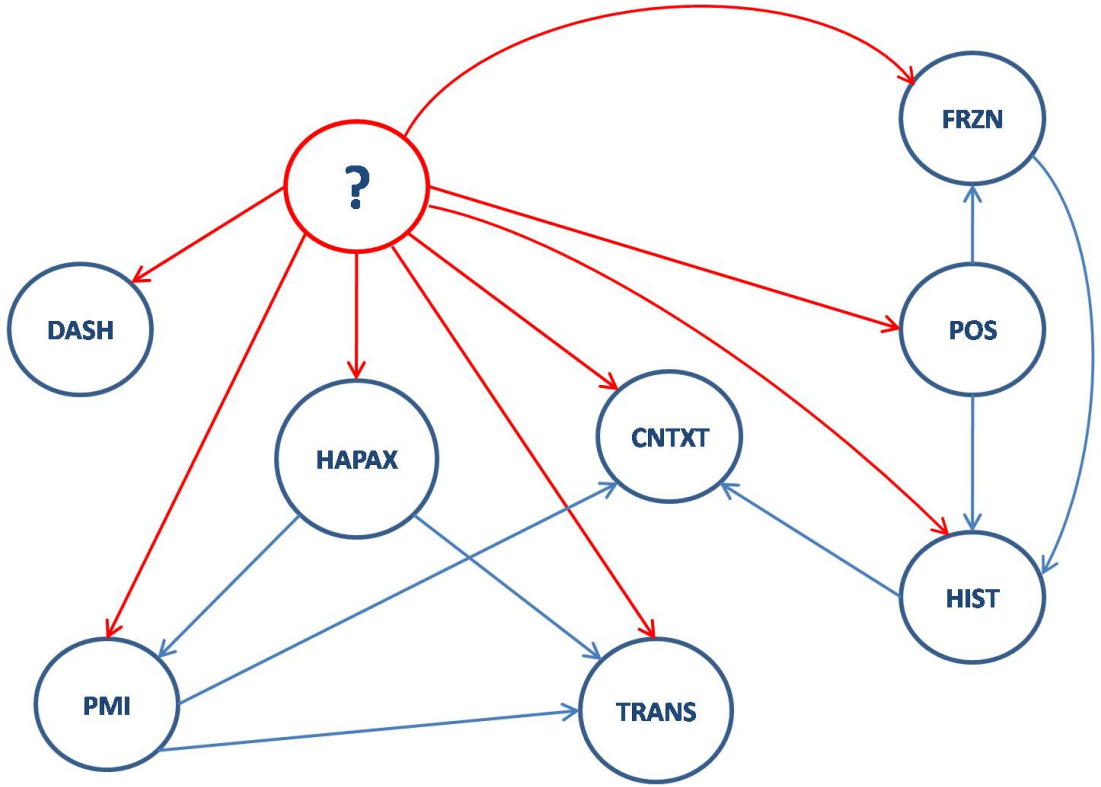
Bayesian Network for MWE identification

Training

- Once the structure of the network is established, the conditional probabilities of each dependency have to be determined.
- We compute the conditional probability tables from our training data using Weka (Hall et al., 2009)
- $P(X \mid X_1, \ldots, X_k)$ for each variable X and all variables X_i, $1 \leq i \leq k$, such that the graph includes an edge from X_i to X (parents of X).

Bayesian Network for MWE identification

Inference



- Using Bayes Rule:

$$P(X_{mwe} \mid X_1, \ldots, X_k) \propto P(X_1, \ldots, X_k \mid X_{mwe}) \times P(X_{mwe})$$

- The prior, $P(X_{mwe}) = 0.41$: the percentage of MWEs in

Training Data Set

- Sample space from which training set is created - aligned and misaligned bi-grams, the output of our previous work:

	Misaligned	Aligned
High PMI	+	-
Low PMI	?	-

- Sizes of the training set:

	MWE	non-MWE	Total
High PMI	300	232	532
Low PMI	50	272	322
Total	350	504	854

- All data, except 50 MWEs with low PMI, were extracted automatically
- Baseline 67%:
  $P(MWE \mid HighPMI) + P(non-MWE \mid LowPMI) = 0.67$

Evaluation

10-fold cross validation

	Accuracy	Precision	Recall	F-score	Error-rate reduction
PMI	66.98%	0.73	0.67	0.67	
BN-auto	71.19%	0.71	0.71	0.71	13%
SVM	74.59%	0.75	0.75	0.75	23%
BN	**76.82%**	**0.77**	**0.77**	**0.77**	**30%**

Evaluation

Hebrew noun-noun constructions

	Accuracy	Precision	Recall	F-score
PMI	71.43%	0.71	0.71	0.71
SVM	77.24%	0.77	0.77	0.77
BN	**77.00%**	**0.77**	**0.77**	**0.77**
AW	80.77%	0.77	0.81	0.79

- AW - Al-Haj and Wintner (2010)

Conclusions

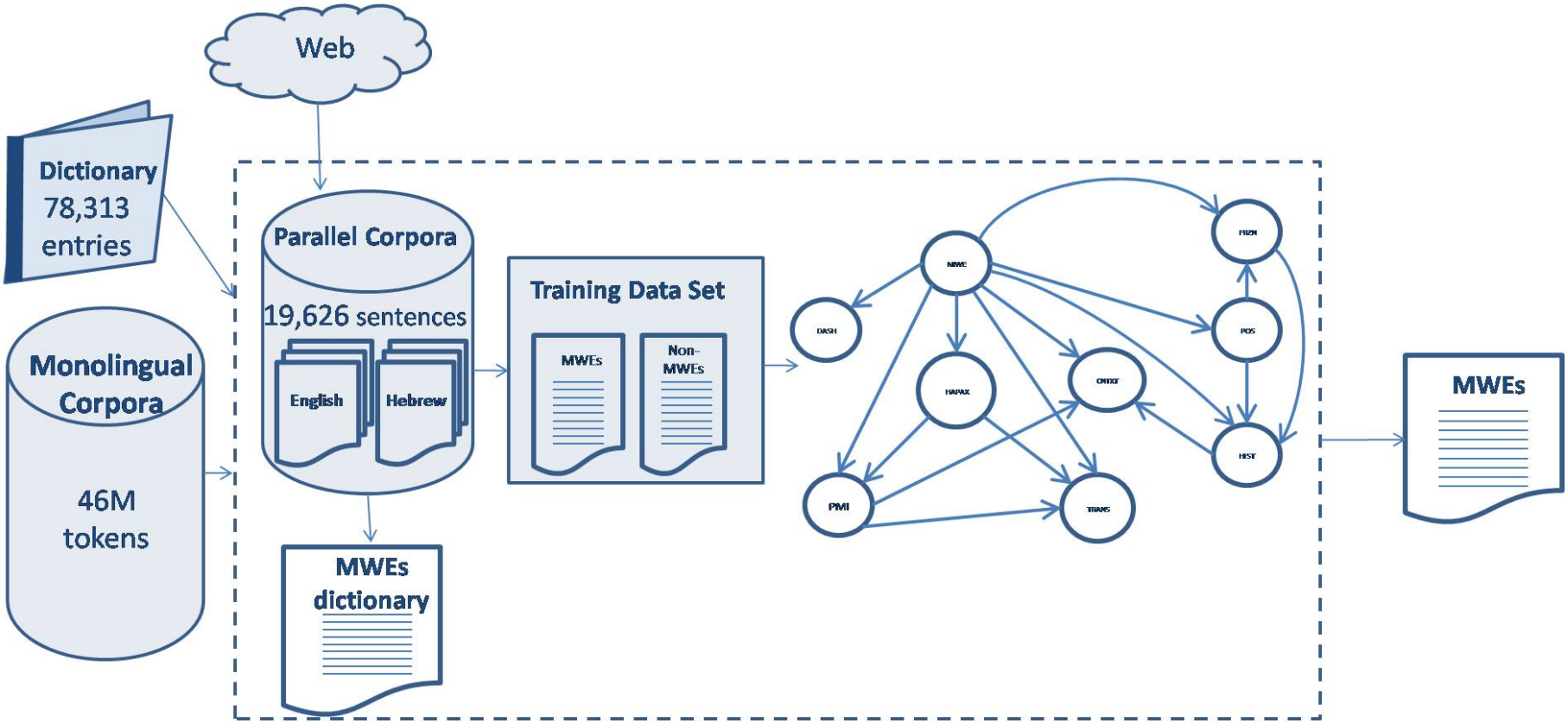
- Features that reflect different facets of irregularities
- Bayesian Network that encodes linguistic knowledge in the features and interdependencies among them
- Automatically extracted training data set

The result is an almost unsupervised, language-independent classification method that can identify MWEs of various lengths, types and constructions.

Future work

- MT – extract dictionary

Questions?


Web
Dictionary
78,313
entries
Monolingual
Corpora
46M
tokens
Parallel Corpora
19,626 sentences
English
Hebrew
MWEs
dictionary
Training Data Set
MWEs
Non-
MWEs
NNVE
FIXN
DASH
POS
HAPAX
CNTXT
HIST
PMI
TRANS
MWEs

Bibliography I

Hassan Al-Haj. Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition. Master's thesis, University of Haifa, February 2010.

Hassan Al-Haj and Shuly Wintner. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 10–18, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL `http://www.aclweb.org/anthology/C10-1002`.

Timothy Baldwin and Takaaki Tanaka. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Colin Bannard. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics, 2007. URL `http://www.aclweb.org/anthology/W/W07/W07-1101`.

Roy Bar-Haim, Khalil Sima'an, and Yoad Winter. Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 39–46, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0706`.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, 2009.

Kenneth. W. Church and Patrick Hanks. Word association norms, mutual information and lexicography (rev). *Computational Linguistics*, 19(1):22–29, 1989.

Antoine Doucet and Helana Ahonen-Myka. Non-contiguous word sequences for information retrieval. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text*, 1(1):29–62, March 2000.

Bibliography II

Afsaneh Fazly and Suzanne Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 337–344, 2006. URL `http://acl.ldc.upenn.edu/E/E06/E06-1043.pdf`.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press, 1998.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://dx.doi.org/10.1145/1656274.1656278`.

Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, March 2008.

Ray Jackendoff. *The Architecture of the Language Faculty*. MIT Press, Cambridge, USA, 1997.

Amit Kirschenbaum and Shuly Wintner. A general method for creating a bilingual transliteration dictionary. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May 2010.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 2004.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Pavel Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, 2008.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Alline Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, 2008.

Yulia Tsvetkov and Shuly Wintner. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1256–1264, Beijing, China, August 2010a. Coling 2010 Organizing Committee. URL `http://www.aclweb.org/anthology/C10-2144`.

Bibliography III

Yulia Tsvetkov and Shuly Wintner. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May 2010b.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. Disambiguating Japanese compound verbs. *Computer Speech & Language*, 19(4):497–512, October 2005.

Tim Van de Cruys and Begoña Villada Moirón. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-1104`.

Sriram Venkatapathy and Aravind Joshi. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, July 2006.

Shailaja Venkatsubramanyan and Jose Perez-Carballo. Multiword expression filtering for building knowledge. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 40–47, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377, 2005.