Multiword Expressions: From Theory to Practicum

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

1. Introduction
2. Robustness in Multiword Expression Identification
   - MWE Token Disambiguation
   - Open-World Token Identification
   - Crosslingual Token Identification
3. Compositionality Prediction
4. Summary

What are Multiword Expressions (MWEs)?

- *Definition:* A **multiword expression** ("MWE") is:
  1. decomposable into multiple simplex words
  2. lexically, phonetically, morphosyntactically, semantically, and/or pragmatically idiosyncratic

Adapted from Baldwin and Kim [2010]

Some Examples

- *East Berlin*, *ad hoc*, *by and large*, *Toy Story*, *kick the bucket*, *part of speech*, *in step*, *ALBA Berlin*, *trip the light fantastic*, *telephone box*, *call (someone) up*, *take a walk*, *do a number on (someone)*, *take advantage (of)*, *pull strings*, *kindle excitement*, *fresh air*, ...

Lexicographic Concept of “Multiword”

- *Heuristic definition:* a lexeme that crosses word boundaries
- Complications with non-segmenting languages (Japanese, Thai, ...) and languages without a pre-existing writing system (Walpiri, Mohawk, ...)
- Also, in English: *houseboat* vs. *house boat*, *trade off* vs. *trade-off* vs. *tradeoff*

Lexical Idiomaticity

- Lexical idiomaticity = one or more of the elements of the MWE does not have a usage outside of MWEs
- Examples of lexical idiomaticity:

  ad hominem, bok choy, a la mode, to and fro

- Complications of lexical idiomaticity:
  - cross-linguistic effects, e.g. *ad* is unmarked in Latin
  - simple lexical occurrence outside of MWEs not sufficient, e.g. *a la mode*

Source(s): Bauer [1983], Trawiński et al. [2008]

Phonetic Idiomaticity

- Phonetic idiomaticity = one or more component elements of the MWE are pronounced in a manner specific to the MWE
- Examples of phonetic idiomaticity:

  cordon bleu, 一期一会 *(ichi-go ichi-e)*

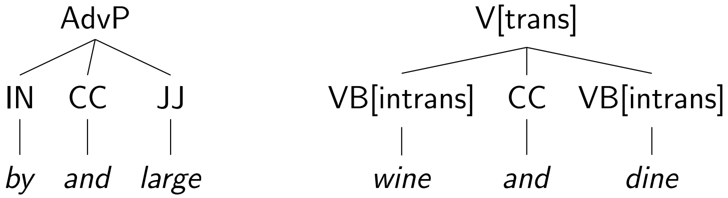
- Also idiosyncratic stress patterns associated with certain MWEs (e.g. *first aid*: Sproat and Liberman [1987])

Morphosyntactic Idiomaticity

- Morphosyntactic idiomaticity = the morphosyntax of the MWE differs from that of its components
- Examples of morphosyntactic idiomaticity:

  cat's cradle, yin hry "evil eye"

- Examples of syntactic idiomaticity:



Source(s): Katz and Postal [2004], Chafe [1968], Bauer [1983], Sag et al. [2002]

Semantic Idiomaticity

- Semantic idiomaticity = the meaning of the MWE is not the simple sum of its parts, in that:
  - there is a mismatch in simplex and MWE semantics for one or more of the components, e.g.

    birds of a feather, blow hot and cold

OR

  - there is extra semantic content in the MWE not encoded in the parts, e.g.

    bus driver (cf. woman driver, backseat driver, valet driver)

Source(s): Katz and Postal [2004], Chafe [1968], Bauer [1983], Sag et al. [2002]

Pragmatic idiomaticity

- Pragmatic idiomaticity = the MWE is associated with a fixed set of situations or a particular context, or with real-world information or expectations about the MWE
- The contexts/real-word information/expectations vary a lot in their generality and also strength:
  - societal norms (e.g. *all aboard*, *gin and tonic*)
  - sub-community norms (e.g. the Monty Python effect)
  - idiolectal norms

Source(s): Kastovsky [1982], Jackendoff [1997], Sag et al. [2002]

Combinational Idiomaticity

- Combinational idiomaticity = a particular combination of words has a high lexical affinity, or preferred lexical configuration relative to alternative phrasings of the same expression, e.g.:

  traffic light, salt and pepper, no worries

- Important to distinguish from "statistical" idiomaticity: statistics is a powerful proxy for capturing combinational idiomaticity, but is not axiomatic

Combinational Idiomaticity

- Closely related to **institutionalisation** = the degree to which a certain expression has come to be used as the preferred way of referring to a given object or concept, among the myriad of different expressions that could plausibly be used to refer to it
- Institutionalisation driven by a myriad of factors, including:
  - phonetics and phonology (e.g. *silly billy*)
  - crosslingual factors (e.g. *willy willy*)
  - sociological factors (e.g. *shock and awe*, *fair play*)
- Important to note that combinational idiomaticity is neither sufficient nor necessary for MWEhood, e.g. *powerful ally*, *armagnac and blackcurrant*

Source(s): Fernando and Flavell [1981], Bauer [1983], Nunberg et al. [1994], Sag et al. [2002]

MWE Markedness

MWE	**Markedness**				
	Lex	**Phon**	**MorSyn**	**Sem**	**Prag**
ad hominem	☑	?	?	?	?
at first	☒	☒	☑	☑	☒
first aid	☒	☑	☒	☑	☑
salt and pepper	☒	☒	☒	☑	☑
good morning	☒	☒	☒	☑	☑
cat's cradle	☒	☒	☑	☑	☑

(Some) NLP Challenges for MWEs

- Robust identification and extraction of MWEs, esp. for languages without MWE resources
- Modelling of semantic compositionality which is faithful to the semantic idiosyncrasies of MWEs
- "Bootstrapping" of MWE analysis for novel languages and MWEs

Talk Outline

Ambiguous MWEs

- Many (verbal) MWEs are ambiguous between a literal and idiomatic interpretation, e.g.:

  Kim kicked the bucket

Source(s): `http://www.flickr.com/photos/paparutzi/165725609/`
`http://www.flickr.com/photos/alimander/5504888605/`

Type-specialised MWE Identification/Disambiguation

- Type-specialised classification (e.g. Hashimoto and Kawahara [2009], Fothergill and Baldwin [2011]):
  - train a classifier for each MWE-type in the corpus, based on token-level annotated data
- **Problems:**
  - classifiers only work on tokens of the type they were trained on
  - requires unrealistically large amounts of annotated data

Robustness Solution v1: Crosstype MWE-token classification

- **Approach:** train a cross-type classifier, and apply it to novel MWE types, based on:
    1. type-level information on the flexibility of the MWE
    2. WSD-style context features

Source(s): Fothergill and Baldwin [2012]

MWE Features

- Idiom features:
  - Lexico-syntactic flexibility of the MWE:

    #kick the pail
    #strike the bucket
    #the bucket was kicked
    #kicking buckets
  - Lexico-semantic features of constituents
- WSD features:
  - semantic vectors (bag of words)
  - selectional preferences
  - local collocations

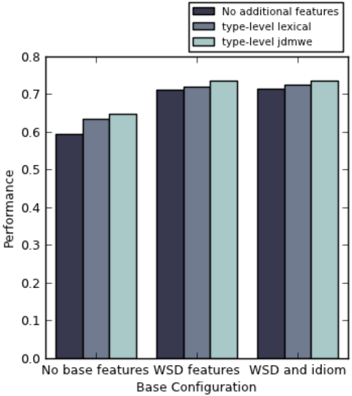
Source(s): Fothergill and Baldwin [2012]

Experiment

- Base experiment on Japanese, and the *OpenMWE* corpus of Japanese idioms (90 MWE-types; $100,000$ tokens: Hashimoto and Kawahara [2009])
- *JDMWE* [Shudo et al., 2011] = a dictionary of thousands of Japanese idioms specifying their relative lexico-syntactic fixedness; compare with type-based features of Fothergill and Baldwin [2011]
- Syntactic features from KNP [Kurohashi and Nagao, 1994]; morphological and lexical semantic features from JUMAN [Kurohashi and Nagao, 1998]
- Experiments based on cross-validation with type-level stratification

Source(s): Fothergill and Baldwin [2012]

Results



Source(s): Fothergill and Baldwin [2012]

Findings

- WSD features lead to surprising accurate; much greater impact than type-level features
- MWE lexicon-based features slightly better than data-driven syntactic features of Fothergill and Baldwin [2011]
- Many instances of violations of the constraints in the MWE lexicon

Source(s): Fothergill and Baldwin [2012]

Robustness Solution v2: MWE-token Identification as Sequence Labelling

- Findings of Fothergill and Baldwin [2012] intriguing, but are predicated on having a pre-existing lexicon of ambiguous MWEs

Source(s): Schneider et al. [2014a], Qu et al. [2015]

Robustness Solution v2: MWE-token Identification as Sequence Labelling

- Findings of Fothergill and Baldwin [2012] intriguing, but are predicated on having a pre-existing lexicon of ambiguous MWEs ... but is MWE identification anything more than sequence labelling?

Source(s): Schneider et al. [2014a], Qu et al. [2015]

Robustness Solution v2: MWE-token Identification as Sequence Labelling

- Findings of Fothergill and Baldwin [2012] intriguing, but are predicated on having a pre-existing lexicon of ambiguous MWEs ... but is MWE identification anything more than sequence labelling?
- **Approach:** train a MWE identification sequence labeller, and apply it to novel data to see whether it can identify novel MWEs

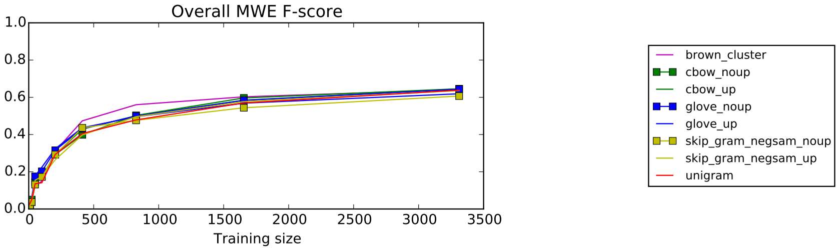
Source(s): Schneider et al. [2014a], Qu et al. [2015]

Experiment

- Base experiment on English, and the MWE corpus of Schneider et al. [2014b] (56K words exhaustively annotated for MWEs)
- Identification based on first-order linear-chain graph transformer [Collobert et al., 2011], optionally using different types of pre-trained word embeddings as input
    - as a by-product of training the model, all words in the training data will end up with fine-tuned type-level representations
- Optionally include lexical features, based on combination of English MWE lexicons

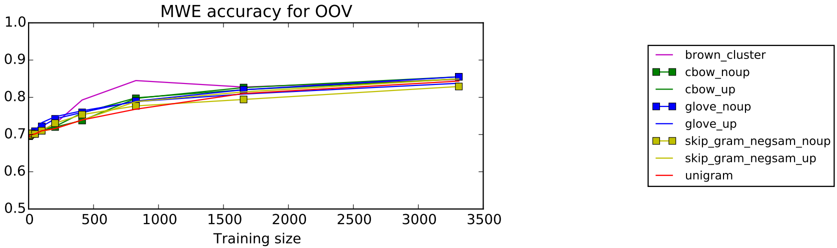
Source(s): Qu et al. [2015]

Results (Overall)



Source(s): Qu et al. [2015]

Results (OOV)



Source(s): Qu et al. [2015]

Findings

- Remarkable ability to classify OOV MWEs
- Lexicons have some impact, but relatively slight (possible to achieve plausible results without lexicons)
- Relatively little difference between the different embeddings

Source(s): Qu et al. [2015]

Robustness Solution v3: MWE-token Identification as Cross-lingual Sequence Labelling

- Impressive results achieved monolingually, but can't always rely on access to token-level annotated MWE data for a given language
- **Approach:**
  1. train a delexicalised POS tagger + dependency parser for a given language and also multilingual word embeddings, based on small amount of parallel data (or just bilingual lexicon)
  2. In the first instance, apply the model to the target language and "read off" the MWEs directly
  3. Add extra constructional features to support construction-level transfer learning

Talk Outline

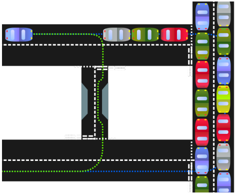
Introduction

- Compositionality prediction = prediction of the relative semantic compositionality ($\in [0,1]$) of a given MWE wrt its component words

climate change
comp = 0.99

silver screen
comp = 0.48

rat run
comp = 0.15



Source(s): Reddy et al. [2011], Schulte im Walde et al. [2013]

Approach v1

- **Hypothesis:** MWE compositionality $\propto$ lexical compositionality under translation
- **Approach:**
  1. look up MWE and also each of the component words in a broad-coverage multilingual dictionary
  2. estimate compositionality based on the combined string similarity between each of the components and the overall MWE, within each of the languages

Source(s): Salehi and Cook [2013]

Approach v2

- **Hypothesis:** MWE compositionality $\propto$ weighted average of distributional similarity between the MWE and each of its components ... possibly combined across a range of languages
- **Approach:**
  1. look up MWE and also each of the component words in a broad-coverage multilingual dictionary
  2. (naively) pre-identify token occurrences of each MWE in a text corpus
  3. calculate the distributional similarity between the MWE and each component word, and combine across the components via weighted mean
  4. combine across languages via the simple arithmetic mean

Source(s): Salehi et al. [2014]

Approach v3

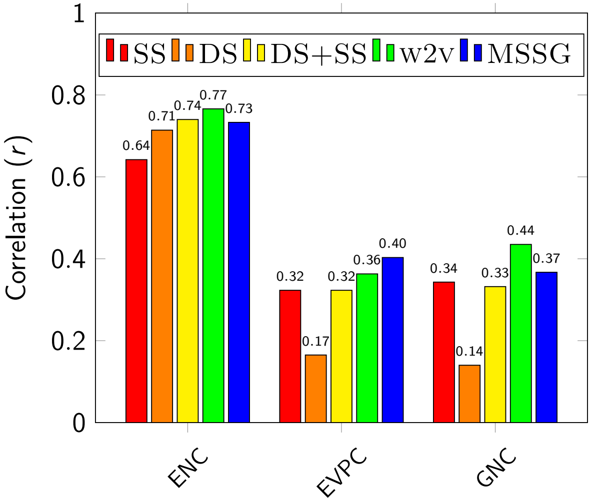
- **Hypothesis:** MWE compositionality $\propto$ weighted average of distributional similarity between the MWE and each of its components ... as estimated based on embedding-based similarity
- **Approach:**
  1. (naively) pre-identify token occurrences of each MWE in a text corpus
  2. pre-train embeddings for the MWE and each component
  3. calculate the distributional similarity between the MWE and each component word based on cosine similarity, and combine across the components via weighted mean
- Experiment with two methods for learning embeddings: WORD2VEC [Mikolov et al., 2013] and MSSG [Neelakantan et al., 2014]

Source(s): Salehi et al. [2015a]

Experiment

- Base experiment on three MWE datasets:
  1. English compound nouns [Reddy et al., 2011]
  2. English verb particle constructions [Bannard, 2006]
  3. German compound nouns [Schulte im Walde et al., 2013]
- As the multilingual dictionary, use PanLex [Baldwin et al., 2010, Kamholz et al., 2014]
- Evaluate based on Pearson's r relative to the gold-standard compositionality judgements

Results


SS
DS
DS+SS
w2v
MSSG
Correlation (r)
1
0.8
0.6
0.4
0.2
0
0.64
0.71
0.74
0.77
0.73
0.32
0.17
0.32
0.36
0.40
0.34
0.14
0.33
0.44
0.37
ENC
EVPC
GNC

Findings

- String similarity over large number of languages (with sub-selection of language) provides a strong unsupervised baseline, and powerful backoff strategy for distributional similarity-based methods
- For tokens which can be identified with suitable frequency in a text corpus, distributional similarity provides a powerful means of predicting compositionality
- In all cases, no language-specific information used by our method and no labelled data required, so applicable to any language/MWE
- Preliminary results to indicate that compositionality predictions can improve MT evaluation [Salehi et al., 2015b]

Talk Outline

Summary

- There's much, much more to MWEs than our old friend *kick the bucket*
- As a complement to "deep dive" work on specific MWEs in specific languages, important to develop automatic language-independent methods for MWE processing
- Increasingly possible to develop methods with the ability to model novel MWEs/MWEs in novel languages ... but still lots more work to do

References

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition, 2010.

Timothy Baldwin, Jonathan Pool, and Susan M. Colowick. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Demo Volume*, pages 37–40, Beijing, China, 2010.

Colin Bannard. *Acquiring Phrasal Lexicons from Corpora*. PhD thesis, University of Edinburgh, UK, 2006.

Laurie Bauer. *English Word-formation*. Cambridge University Press, Cambridge, UK, 1983.

Wallace L. Chafe. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language*, 4:109–127, 1968.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Chitra Fernando and Roger Flavell. *On idioms*. Exeter: University of Exeter, 1981.

References

Richard Fothergill and Timothy Baldwin. Fleshing it out: A supervised approach to MWE-token and MWE-type classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 911–919, Chiang Mai, Thailand, 2011.

Richard Fothergill and Timothy Baldwin. Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 100–104, Montreal, Canada, 2012.

Chikara Hashimoto and Daisuke Kawahara. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43:355–384, 2009.

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

David Kamholz, Jonathan Pool, and Susan Colowick. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3145–3150, Reykjavik, Iceland, 2014.

Dieter Kastovsky. *Wortbildung und Semantik*. Bagel/Francke, Dusseldorf, Germany, 1982. (in German).

References

Jerrold J. Katz and Paul M. Postal. Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report (70), MIT Research Laboratory of Electronics*, pages 275–282. MIT Press, 2004.

Sadao Kurohashi and Makoto Nagao. KN parser: Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*, Nara, Japan, 1994.

Sadao Kurohashi and Makoto Nagao. *Nihongo keitai-kaiseki sisutemu JUMAN* [Japanese morphological analysis system JUMAN] version 3.5. Technical report, Kyoto University, 1998. (in Japanese).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA, 2013.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1059–1069, Doha, Qatar, 2014.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. Idioms. *Language*, 70:491–538, 1994.

References

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider, and Timothy Baldwin. Big data small data, in domain out-of domain, known word unknown word: The impact of word representations on sequence labelling tasks. In *Proceedings of the 19th Conference on Natural Language Learning (CoNLL-2015)*, pages 83–93, Beijing, China, 2015.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 210–218, Chiang Mai, Thailand, 2011.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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, 2002.

Bahar Salehi and Paul Cook. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 266–275, Atlanta, USA, 2013.

References

Bahar Salehi, Paul Cook, and Timothy Baldwin. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*, pages 472–481, Gothenburg, Sweden, 2014.

Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2015)*, pages 977–983, Denver, USA, 2015a.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the NAACL HLT 2015 Workshop on Multiword Expressions*, pages 54–59, Denver, USA, 2015b.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, 2014a. URL `http://www.transacl.org/wp-content/uploads/2014/04/51.pdf`.

References

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 455–461, Reykjavík, Iceland, 2014b. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf`.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 255–265, Atlanta, USA, 2013.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 161–170, Portland, USA, 2011.

Richard W. Sproat and Mark Y. Liberman. Toward treating English nominals correctly. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, USA, 1987.

References

Beata Trawiński, Manfred Sailer, Jan-Philipp Soehn, Lothar Lemnitzer, and Frank Richter. Cranberry expressions in English and in German. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 35–38, Marrakech, Morocco, 2008.