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 (Walpirti, 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:
  
  \textit{cordon bleu, 一期一会} (ichi-go ichi-e)

- Also idiosyncratic stress patterns associated with certain MWEs (e.g. \textit{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:

  \[
  \begin{align*}
  \text{AdvP} & \quad \text{V[trans]} \\
  \text{IN} & \quad \text{VB[intrans]} \quad \text{CC} \quad \text{VB[intrans]} \\
  \text{by} & \quad \text{and} \quad \text{large} \quad \text{wine} \quad \text{and} \quad \text{dine}
  \end{align*}
  \]

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.:
  
  \textit{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

<table>
<thead>
<tr>
<th>MWE</th>
<th>Lex</th>
<th>Phon</th>
<th>MorSyn</th>
<th>Sem</th>
<th>Prag</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ad hominem</em></td>
<td>✓</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><em>at first</em></td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td><em>first aid</em></td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><em>salt and pepper</em></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><em>good morning</em></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><em>cat’s cradle</em></td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
(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

1. Introduction

2. Robustness in Multiword Expression Identification
   - MWE Token Disambiguation
   - Open-World Token Identification
   - Crosslingual Token Identification

3. Compositionality Prediction

4. Summary
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)

MWE accuracy for 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
Introduction

Robustness in Multiword Expression Identification
- MWE Token Disambiguation
- Open-World Token Identification
- Crosslingual Token Identification

Compositionality Prediction

Summary
Introduction

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

\begin{itemize}
  \item \textit{climate change} \hspace{1cm} \textit{silver screen} \hspace{1cm} \textit{rat run}
  \item \text{comp} = 0.99 \hspace{1cm} \text{comp} = 0.48 \hspace{1cm} \text{comp} = 0.15
\end{itemize}

\textbf{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]
  - 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

![Graph showing correlation coefficients]

- ENC:
  - SS: 0.71
  - DS: 0.74
  - DS+SS: 0.77
  - w2v: 0.73
- EVPC:
  - SS: 0.64
  - DS: 0.32
  - DS+SS: 0.36
  - w2v: 0.40
  - MSS: 0.34
- GNC:
  - SS: 0.17
  - DS: 0.14
  - DS+SS: 0.33
  - w2v: 0.44
  - MSS: 0.37
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

1. Introduction

2. Robustness in Multiword Expression Identification
   - MWE Token Disambiguation
   - Open-World Token Identification
   - Crosslingual Token Identification

3. Compositionality Prediction

4. Summary
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 in novel languages ... but still lots more work to do
References


References


References


References


References


References


References