Bilingual Lexicon Extraction at the Morpheme Level Using Distributional Analysis

Amir HAZEM and Béatrice DAILLE

LINA, University of Nantes, France
10th edition of the Language Resources and Evaluation Conference (LREC 2016)

23-28 May 2016, Portorož (Slovenia)

Outline

Outline

Outline

Outline

Outline

Outline

Introduction

- bilingual single word terms (SWT) extraction –> usually based on distributional methods
- no consideration of the compositional property of SWTs
- many SWTs are compositional (composed of roots and affixes)
- compositionality can be very useful to match translational pairs (especially for infrequent terms where distributional methods often fail)

Introduction (Example)

- to translate *xenograft* from English to French
  - root *xeno* is aligned with *xéno*
  - lexeme *graft* is aligned with *greffe*
- this results in the translation *xénogreffe* (based on compositionality)

Introduction

- we experiment several distributional modellings at the morpheme level
- we apply compositional translation to a subset of French and English compounds
- we show promising results using distributional analysis at the root and affix levels
- we also show that the adapted approach significantly improve bilingual lexicon extraction from comparable corpora compared to the approach at the word level

Related work

- Distributional Semantic Models (DSM) are used in many nlp tasks [Guevara, 2010]
  - bilingual terminology extraction from comparable corpora (SWT's) [Rapp, 1999, Laroche and Langlais, 2010, Morin and Hazem, 2014]
- one of the main problems of DSMs is data sparseness
- derivational morphology property of SWT should help

Related work

- compositional methods (originally developed for phrases) were successfully applied:
  - [Delpech et al., 2012] translate morphologically constructed terms (exploite a manually constructed translation list at the morpheme-level)
  - [Guevara, 2010, Lazaridou et al., 2013] improve the quality of monolingual neighbor acquisition

Related work

- [Lazaridou et al., 2013] adapted compositional methods to the task of deriving the distributional meaning of morphologically complex words from their parts
- they explored the application of compositional distributional semantic models (cDSM) to derivational morphology by adapting several composition methods
  - multiplicative model, additive model, full additive model, etc.
  - applyed the lexical function model (*lexfunc*) [Baroni and Zamparelli, 2010] where the distributional representation of one element in a composition is not a vector but a function
  - also used the DSM at the stem level as a baseline
- our approach is inspired from [Lazaridou et al., 2013] and uses the additive model to combine several distributional modellings at the morpheme level

Various forms of compounds

- closed compounds [Macherey et al., 2011]: written as single words (*toolbar*)
- open compounds: space-separated but form a unit of meaning (*operating system*)
- we only deal with closed compounds (including hyphen-separated)
- major kinds of compounds are native and neoclassical
  - the first kind includes only native elements (not borrowed from another language suh as *parrot + fish = parrotfish*)
  - the second kind, neoclassical compounding (combines some elements of Greek or Latin etymological origin, such as *hydro + logy = hydrology*

Various forms of compounds

- neoclassical elements are not considered as lexical units (never occur independently e.g., *biology* [Amiot and Dal, 2008, Namer, 2009])
- each language may assimilate its borrowed neoclassical elements phonologically (but not totally) [Lüdeling, 2006]
- a Greek or Latin word undergoes a minimal adaptation before being adopted by a host language (For example, both Fr: *pathie* and En: *pathy* were borrowed from the Greek word *pathos*
- prefixed words cannot be called compounds in the strict sense of the term (prefixes are not independent lexical units)

Various forms of compounds

- some prefixes are very close to the neoclassical roots, compare prefix *bi-* with neoclassical root *uni-* according to [Béchade, 1992]
- the difference is in their origin (neoclassical roots come from Latin or Greek content words, whereas prefixes come from function words) and the period when they entered into usage (prefixes entered earlier)
- we focus on neoclassical compounds and prefixed words, and compounds such as hidden compounds which are at the border between native and neoclassical compounds
  - including elements that are not purely neoclassical elements but look like them, such as the element *radio* in *radiology*

Bilingual morpheme extraction (proposed approach)

- extract for each source morpheme (root or affix) its corresponding translation in a target language
- adapt a distributional method to the morpheme level

Bilingual morpheme extraction (proposed approach)

1. split each SWT into roots or affixes and lexemes (*abnormal* –> *ab* and *normal*)
   - many existing splitting tools DeriF (Namer 2003)[Namer, 2009] for French, language independent such as Koehn and Knight (2003)[Koehn and Knight, 2003] algorithm or COMPOST [Loginova Clouet and Daille, 2014]
2. build context vector of affix or root (contains lexemes (*ab*: *normal*))
3. weight each lemma or lexeme (association measure)
4. translate each source context vector into the target language using a bilingual dictionary
5. apply a similarity measure between each translated source context vector and all the target vectors
6. rank the translation candidates according to their similarity scores

four variants of the proposed approach (*lexem*)

- add the single term (lemma) to the context vector of the morpheme (noted *lem*)
  - add *abnormal* to *ab*
- add lexeme and the single term (lemma) to the context vector of the morpheme (noted *lexem* + *lem*)
  - add *normal* and *abnormal* to *ab*
- add context words of the lexeme to the context vector of the morpheme (noted *Vect*(*lexem*)
  - add context words of *normal* to the prefix *ab*
- add the single term context words to the context vector of the morpheme (noted *Vect*(*lem*))
  - *abnormal* context words will be added to the context vector of the prefix *ab*

Experiments and results

- two sets of experiments have been conducted
  - bilingual morphemes (roots and affixes) extraction
  - bilingual terminology extraction using the extracted bilingual morphemes

Experiments and results

- two French-English comparable corpora
  - breast cancer corpus of 500k words
  - wind energy corpus of 400k words
- two reference lists (selected French-English morpheme pairs from an existing list of 892 entries)
  - 176 morpheme translations from the breast cancer corpus
  - 80 morpheme translations from the wind energy corpus
- one dictionary: we used the French/English ELRA-M0033 resource (a general language dictionary which contains only a few terms related to specialised domains)

Experiments and results

- impact of bilingual morpheme extraction on the bilingual terminology extraction task
  - an additional list of morphologically derived terms on the breast cancer corpus has been built (32 translation pairs)

Experimental setup

- distributional method
  - association measure –> log-likelihood [Dunning, 1993]
  - similarity measure –> weighted jaccard index [Grefenstette, 1994]
  - a 7-window context vectors size
- evaluation
  - precision at P1, P5, P10 and accuracy (Acc.)
  - mean average precision *MAP* [Manning et al., 2008]

Bilingual morpheme extraction results on the breast cancer corpus (EN-FR)

	P1	P5	P10	Acc.	MAP
lexem	16.9	35.0	38.0	39.7	23.9
lem	**30.9**	32.7	32.7	32.7	31.8
lexem + *lem*	27.4	**40.3**	**43.2**	43.2	32.7
Vect(*lexem*)	21.0	33.3	39.1	**61.9**	27.8
Vect(*lem*)	28.6	38.5	42.1	60.8	**34.1**

Table : Results (%) of morphemes translation for the breast cancer corpus (en-fr)

Bilingual morpheme extraction results on the breast cancer corpus (FR-EN)

	P1	P5	P10	Acc.	MAP
lexem	19.2	36.2	38.5	41.5	26.2
lem	30.9	32.7	32.7	32.7	31.8
lexem + *lem*	**33.9**	**43.8**	**45.0**	45.0	**38.5**
Vect(*lexem*)	25.1	35.0	38.0	**68.4**	29.8
Vect(*lem*)	29.8	40.3	44.4	67.2	35.2

Table : Results (%) of morphemes translation for the breast cancer corpus (fr-en)

Bilingual morpheme extraction results on the wind energy corpus (EN-FR)

	P1	P5	P10	Acc.	MAP
lexem	23.7	36.2	38.7	38.7	29.1
lem	**38.7**	40.0	40.0	40.0	39.2
lexem + *lem*	36.2	**43.7**	45.0	45.0	**39.4**
Vect(*lexem*)	25.0	38.7	43.7	63.7	31.0
Vect(*lem*)	31.2	**43.7**	**47.5**	**65.0**	37.2

Table : Results (%) of morphemes translation for the wind energy corpus (en-fr)

Bilingual morpheme extraction results on the wind energy corpus (FR-EN)

	P1	P5	P10	Acc.	MAP
lexem	28.7	36.2	37.5	37.5	31.3
lem	**40.0**	40.0	40.0	40.0	**40.0**
lexem + *lem*	36.2	**43.7**	**45.0**	45.0	39.1
Vect(*lexem*)	22.5	37.5	42.5	**67.5**	29.5
Vect(*lem*)	25.0	38.5	43.7	66.2	31.6

Table : Results (%) of morphemes translation for the wind energy corpus (fr-en)

Bilingual terminology extraction results

	P1	P5	P10	Acc.	MAP
DistApp	6.25	12.5	18.7	**50.0**	10.7
lexem	21.8	21.8	21.8	21.8	21.9
lem	**34.3**	**37.5**	**37.5**	37.5	**35.9**
lexem + *lem*	18.7	25.0	25.0	25.0	20.4
Vect(*lexem*)	18.7	25.0	25.0	25.0	20.4
Vect(*lem*)	18.7	25.0	25.0	25.0	20.4

Table : Results (%) of bilingual term extraction for the breast cancer corpus (en-fr)

Conclusion and future work

- a new method based on distributional semantics to automatically build bilingual translations at the morpheme-level
- to our knowledge, this work is the first evaluation of such a task
- additional experiments for other languages than English and French are certainly needed
- we can at least conclude that morphological analysis associated to distributional semantics is appropriate for bilingual morphemes alignment as well as for bilingual terminology extraction from comparable corpora

Conclusion and future work

- our experiments have been conducted with a reference segmentation, that is a manual segmentation
- we foresee in our next experiments to use as input segmentations provided by a splitting tool
- we need to investigate how distributional analysis at the morpheme level deals with erroneous splitting
- [Clouet et al., 2015] compared manual and automatic segmentations for a translation task using a compositional translation
  - they showed that the results are similar when a precision-oriented segmentation was chosen for the automatic splitting
- we hope to reach the same conclusion with automatically built bilingual morpheme translations

Thank you for your attention

Thank you for your attention

Amiot, D. and Dal, G. (2008).
La composition néoclassique en français et l'ordre des constituants.
La composition dans les langues, Artois Presses Université, pages 89–113.

Baroni, M. and Zamparelli, R. (2010).
Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.
In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1183–1193.

Béchade, H.-D. (1992).
Phonétique et morphologie du français moderne et contemporain.
Presses Universitaires de France.

Clouet, E., Harastani, R., Daille, B., and Morin, E. (2015).
Compositional translation of single-word complex terms using multilingual splittin.

Terminology. Special Issue: Terminology across languages and domains, 21(2).


Delpech, E., Daille, B., Morin, E., and Lemaire, C. (2012).
Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking.
In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762.


Dunning, T. (1993).
Accurate Methods for the Statistics of Surprise and Coincidence.
Computational Linguistics, 19(1):61–74.


Grefenstette, G. (1994).
Explorations in Automatic Thesaurus Discovery.
Kluwer Academic Publisher, Boston, MA, USA.


Guevara, E. (2010).
A regression model of adjective-noun compositionality in distributional semantics.

In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003).
Empirical methods for compound splitting.
In *Proceedings of EAC 2003*, Budapest, Hungary.

Laroche, A. and Langlais, P. (2010).
Revisiting context-based projection methods for term-translation spotting in comparable corpora.
In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013).
Compositional-ly derived representations of morphologically complex words in distributional semantics.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1517–1526.

Loginova Clouet, E. and Daille, B. (2014).
Splitting of Compound Terms in non-Prototypical Compounding Languages.
In *Workshop on Computational Approaches to Compound Analysis, COLING 2014*, pages 11 – 19, Dublin, Ireland.

Lüdeling, A. (2006).
Neoclassical word-formation.
In *Keith Brown (ed) Encyclopedia of Language and Linguistics, 2nd Edition*, Oxford, Elsevier.

Macherey, K., Dai, A., Talbot, D., Popat, A., and Och, F. (2011).
Language-independent compound splitting with morphological operations.
In *Proceedings of ACL 2011*, pages 1395–1404, Portland, Oregon.

Manning, D. C., Raghavan, P., and Schütze, H. (2008).
Introduction to information retrieval.
Cambridge University Press.

Morin, E. and Hazem, A. (2014).

Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1284–1293.


Namer, F. (2009).
Morphologie, lexique et traitement automatique des langues.
Lavoisier, Paris.


Rapp, R. (1999).
Automatic Identification of Word Translations from Unrelated English and German Corpora.
In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.