Word Embedding Approach for Synonym Extraction of Multi-Word Terms

Amir Hazem and Béatrice DAILLE

LS2N, Université de Nantes, France
11th edition of the Language Resources and Evaluation Conference (LREC)

7-12 May, 2018 Miyazaki, Japan

Outline

Outline

1. Context
2. Related work
   - Synonyms acquisition
   - Multi-word terms (MWTs) and their synonyms

Outline

Outline

Outline

Outline

Multi-Word Terms (MWTs) Synonyms Acquisition

- relatively **new**, under represented topic
- **challenging** (MWT's synonyms and semantically related terms)
  - MWT's synonyms are single word terms (SWTs)
  - MWT's synonyms are MWT's of different lengths
- **several** researches addressed synonym extraction of **SWTs**
- **few** of them dealt with MWTs and fewer, unless none while **MWT's synonyms are of variable lengths**

Synonyms acquisition (**SWT**)

Synonyms acquisition has mainly concerned SWTs:

- lexicon-based approaches [Blondel and Senellart, 2002]
- multilingual approaches [Wu and Zhou, 2003,
  van der Plas and Tiedemann, 2006, Andrade et al., 2013]
- distributional approaches [Lin, 1998, Hagiwara, 2008]
- distributed approaches (word embeddings) [Mikolov et al., 2013],
  etc.

Synonyms acquisition (**MWT**)

MWTs and their synonyms –> useful in many applications:

- word sense disambiguation
- machine translation
- information retrieval
- text simplification, etc.

Compositionality and Synonymic Variants

Compositionality

Compositionality means that the whole meaning can be deduced from the meaning of its components and the syntactic rule by which they are combined [Partee et al., 1990]

Synonymic variants of multi-words exhibit multiple phenomena:

- MWT synonyms of the same length: ***wind turbine*/*wind machine***
- MWT synonyms of variable length: ***wind farm*/*wind power plant***
- non compositional MWT synonyms: ***pole tower*/*mast***

Synonyms acquisition of MWTs

Main approaches deal with synonyms of MWTs that are:

- **compositional**
- often of the **same length**

Compositionality by substituting parts of the MWT

- synonyms provided by a dictionary [Hamon and Nazarenko, 2001]
- synonyms and semantically related terms provided by distributional analysis [Hazem and Daille, 2014]

Compositional Approach [Hamon and Nazarenko, 2001]

- substitutes one of the components of the MWT by one of its synonyms
  - R_1: $T_1 = T_2 \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$
  - R_2: $E_1 = E_2 \wedge syn(T_1, T_2) \supset syn(CCT_1, CCT_2)$
  - R_3: $syn(T_1, T_2) \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$
- validates the MWT's synonym if found in the corpus

Compositional Approach ([Hamon and Nazarenko, 2001])

- example: ***collecteur général*** 'general collector' synonym?
  - dictionary provides several synonyms of ***général***:
    - ***habituel***
    - ***ordinaire***
    - ***commun***, ...
  - ***collecteur commun*** 'common collector' is the correct synonym
  - it occcurs in the wind energy corpus

Compositional Approach ([Hamon and Nazarenko, 2001])

- example: ***collecteur général*** 'general collector' synonym?
  - dictionary provides several synonyms of ***général***:
    - ***habituel***
    - ***ordinaire***
    - ***commun***, ...
  - ***collecteur commun*** 'common collector' is the correct synonym
  - it occcurs in the wind energy corpus
- based on a dictionary of synonyms
- remains resource dependent

Semi-Compositional Approach ([Hazem and Daille, 2014])

- extended approach based on distributional analysis
- based on the principle of compositionality of MWTs
- main difference lies on the nature of the substitutions
- no longer constrained by the sole relation of synonymy
- generalized substitutions to semantically related terms
  - $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
  - $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$
  - remove R_3, less productive and reliable rule ([Hamon and Nazarenko, 2001])

Semi-Compositional Approach ([Hazem and Daille, 2014])

Example: ***énergie renouvelable*** 'renewable energy':

- extract each part of the MWT
- find the semantically related words of ***énergie*** 'energy' and/or ***renouvelable*** 'renewable' with distributional methods
- filter all expressions using monolingual specialized corpora

Limitations

- resource dependent (dictionary)
- fixed length

Proposed approach (Motivations)

- word embbedings
- efficiently represents phrases, sentences, paragraphs and more generally, pieces of texts of any length
  [Mitchell and Lapata, 2010, Mikolov et al., 2013,
  Socher et al., 2011, Mikolov et al., 2013, Le and Mikolov, 2014,
  Kalchbrenner et al., 2014, Kiros et al., 2015, Wieting et al., 2016,
  Arora et al., 2017, Hazem et al., 2017]
- additive property –> key information for representing phrases by extension MWTs and there synonyms or quasi-synonyms

Proposed 2 Word Embeddings Approaches

- ***Semi-compositional word embeddings***
  - follows the principle of the semi-compositional approach
  - it mainly differs in the procedure of extracting SWTs synonyms or semantically related terms
- ***Full-compositional word embeddings***
  - inspired by phrases representation by an element-wise sum of the word embeddings ([Mikolov et al., 2013])
  - also sentence representation performed by an element wise addition of word embeddings of its parts
    [Wieting et al., 2016, Arora et al., 2017, Hazem et al., 2017]
  - adaptation to MWT synonyms extraction

Semi-Compositional Word Embeddings

- based on the composition of the elements of MWTs
- can be considered as a variant of the semi-compositional approach introduced in [Hazem and Daille, 2014]
- use distributed models: the Skip-gram model and the continuous bag-of-word model (CBOW) ([Mikolov et al., 2013])

Full-Compositional Word Embeddings

- aims at extracting MWTs synonyms of any length
- provides a joint representation for all the MWTs which facilitates MWTs comparison
- MWTs are represented by a single embedding vector (Additive propoerty)
- cosine similarity measure is applied to extract MWTs synonyms
- only n-grams present in the corpus are kept

Corpora

- Fr/En wind energy corpus of 400,000 words
  - craweled from the web
- Fr/En breast cancer corpus of 500,000 words
  - Istex portal using in-domain keywords

Reference Lists

- built from various terminological resources
- only the resources that list synonymic terms in their terminological records have been examined
  - French Terminalf portal
  - English glossary and Termium portal
- wind energy –> 34 French synonyms/20 English synonyms
- breast cancer –> 20 French synonyms/16 English synonyms
- extra list of MWT synonyms of variable lengths
  - 10 French synonyms/9 English synonyms (wind energy corpus)
  - difficult to have lists of large size

Examples of MWTs of same lengths

English term synonyms

wind turbine	wind machine
power supply	energy supply
power plant	electricity plant
savonius model	savonius type
energy output	energy production

French term synonyms

énergie renouvelable	énergie durable
centrale électrique	centrale éolienne
unité de stockage	dispositif de stoskage
arbre primaire	arbre lent
force du vent	vitesse du vent

Table: Examples of English/French synonyms and quasi-synonyms of MWTs recorded in terminology banks of the wind energy domain

Examples of MWTs of variable lengths

English term synonyms

wind power plant	wind farm
wind turbine	windmill
pole tower	mast
reference area	rotor swept area
savonius rotor	vertical axis wind turbine
wind generator	aerogenerator

French term synonyms

alternateur	générateur synchrone
moulin à hélice	éolienne à axe horizontal
panémone	éolienne à axe vertical
implantation	parc éolien
éolienne axe vertical	rotor de darrieus

Table: Examples of English/French synonyms and quasi-synonyms of MWTs recorded in terminology banks of the wind energy domain

Table: Results (MAP%) on the wind energy corpus

Method	French	English
Hamon&Nazarenko	0.25	3.63
Phrase-based (Mikolov)	4.56	6.78

Table: Results (MAP%) on the wind energy corpus

Method	French	English
Hamon&Nazarenko	0.25	3.63
Phrase-based (Mikolov)	4.56	6.78
Semi-Comp (IM-COS)	27.4	32.6
Semi-Comp (DOR-COS)	26.8	27.2
Semi-Comp (LL-JAC)	31.4	36.1

Table: Results (MAP%) on the wind energy corpus

Method	French	English
Hamon&Nazarenko	0.25	3.63
Phrase-based (Mikolov)	4.56	6.78
Semi-Comp (IM-COS)	27.4	32.6
Semi-Comp (DOR-COS)	26.8	27.2
Semi-Comp (LL-JAC)	31.4	36.1
Semi-Comp (SG50)	30.9	50.3
Semi-Comp (SG100)	**34.9**	55.9
Semi-Comp (SG200)	34.8	52.7
Semi-Comp (CBOW50)	23.0	49.0
Semi-Comp (CBOW100)	23.7	49.4
Semi-Comp (CBOW200)	23.8	49.4

Table: Results (MAP%) on the wind energy corpus

Method	French	English
Hamon&Nazarenko	0.25	3.63
Phrase-based (Mikolov)	4.56	6.78
Semi-Comp (IM-COS)	27.4	32.6
Semi-Comp (DOR-COS)	26.8	27.2
Semi-Comp (LL-JAC)	31.4	36.1
Semi-Comp (SG50)	30.9	50.3
Semi-Comp (SG100)	**34.9**	55.9
Semi-Comp (SG200)	34.8	52.7
Semi-Comp (CBOW50)	23.0	49.0
Semi-Comp (CBOW100)	23.7	49.4
Semi-Comp (CBOW200)	23.8	49.4
Full-Comp (SG100)	27.3	57.8
Full-Comp (SG200)	28.9	**58.4**
Full-Comp (SG300)	28.5	55.3
Full-Comp (CBOW50)	22.6	47.0
Full-Comp (CBOW100)	20.1	45.1
Full-Comp (CBOW200)	21.6	44.5

Table: Results (MAP%) on the breast cancer corpus

Method	French	English
Hamon&Nazarenko	4.92	7.03
Phrase-based (Mikolov)	8.37	9.12

Table: Results (MAP%) on the breast cancer corpus

Method	French	English
Hamon&Nazarenko	4.92	7.03
Phrase-based (Mikolov)	8.37	9.12
Semi-Comp (IM-COS)	19.9	12.6
Semi-Comp (LO-COS)	27.1	11.0
Semi-Comp (LL-JAC)	13.9	13.3

Table: Results (MAP%) on the breast cancer corpus

Method	French	English
Hamon&Nazarenko	4.92	7.03
Phrase-based (Mikolov)	8.37	9.12
Semi-Comp (IM-COS)	19.9	12.6
Semi-Comp (LO-COS)	27.1	11.0
Semi-Comp (LL-JAC)	13.9	13.3
Semi-Comp (SG50)	32.1	15.0
Semi-Comp (SG100)	**32.2**	15.2
Semi-Comp (SG300)	27.9	9.60
Semi-Comp (CBOW50)	29.1	15.1
Semi-Comp (CBOW100)	29.2	15.3
Semi-Comp (CBOW300)	29.4	15.8

Table: Results (MAP%) on the breast cancer corpus

Method	French	English
Hamon&Nazarenko	4.92	7.03
Phrase-based (Mikolov)	8.37	9.12
Semi-Comp (IM-COS)	19.9	12.6
Semi-Comp (LO-COS)	27.1	11.0
Semi-Comp (LL-JAC)	13.9	13.3
Semi-Comp (SG50)	32.1	15.0
Semi-Comp (SG100)	**32.2**	15.2
Semi-Comp (SG300)	27.9	9.60
Semi-Comp (CBOW50)	29.1	15.1
Semi-Comp (CBOW100)	29.2	15.3
Semi-Comp (CBOW300)	29.4	15.8
Full-Comp (SG100)	25.6	17.4
Full-Comp (SG200)	28.0	**18.9**
Full-Comp (SG300)	30.5	16.0
Full-Comp (CBOW100)	24.9	10.6
Full-Comp (CBOW200)	24.9	11.6
Full-Comp (CBOW300)	25.0	10.5

Full-Compositional (Lists of variable lengths)

- extra experiment only on synonyms of variable lengths
- wind energy corpus for French and English
- MAP score of 10.2% and a recall of 66.6% for English
- MAP score of 4.46% of MAP and a recall of 40% for French
- state of art and Semi-comp proposed approaches can't be applied for this experiment (don't deal with MWTs length variability)
- results of Full-Comp approach are still low
- offers an alternative to pairs of MWTs synonyms that have different lengths
- still much to do...

Conclusion and Future Work

- proposed different word embeddings approaches for synonyms extraction of MWTs
- word embeddings with compositionality and additive composition improve the results comparing to baseline approaches (Same length)
- an alternative to MWTs synonyms of variable length
  - first attemp
  - resuls still low
  - no specific filtering process
- open questions:
  - productivity
  - how to fix length variability
  - how to manage repetition in the MWTs
  - predicting the compositionality based on methods borrowed from Multiword Expressions approaches [Salehi et al, NAACL 2015]

Full-Comp Error Analysis

- wind power station
  - power plant
  - wind power
  - power
  - plant
  - wind
  - wind turbine
  - **wind farm**
  - wind energy
  - wind speed
  - wind farm development
  - wind generator
  - offshore wind farm
  - wind turbine noise
  - wind turbine sound

Full-Comp Error Analysis

- darrieus rotor
  - rotor
  - darrieus
  - equip
  - untwisted
  - stator
  - wrig
  - alignment
  - **vertical axis wind turbine**
  - align
  - vertical axis wind
  - rotate
  - leeward
  - tip
  - blade

Thank you.

The implementation of the Semi-compositional and the Full-compositional approaches will be soon available at `https://github.com/hazemAmir/FullComp.git`

Thank you.

The implementation of the Semi-compositional and the Full-compositional approaches will be soon available at
`https://github.com/hazemAmir/FullComp.git`

Andrade, D., Tsuchida, M., Onishi, T., and Ishikawa, K. (2013).
Synonym acquisition using bilingual comparable corpora.
In *International Joint Conference on Natural Language Processing (IJCNLP'13)*, Nagoya, Japan.

Arora, S., Yingyu, L., and Tengyu, M. (2017).
A simple but tough to beat baseline for sentence embeddings.
In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

Blondel, V. D. and Senellart, P. (2002).
Automatic extraction of synonyms in a dictionary.

Hagiwara, M. (2008).
A supervised learning approach to automatic synonym identification based on distributional features.
In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 1–6, Columbus, Ohio. Association for Computational Linguistics.

Hamon, T. and Nazarenko, A. (2001).

Detection of synonymy links between terms: experiment and results.
In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.


Hazem, A. and Daille, B. (2014).
Semi-compositional method for synonym extraction of multi-word terms.
In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).


Hazem, A., el amel Boussaha, B., and Hernandez, N. (2017).
Mappsent: a textual mapping approach for question-to-question similarity.
Recent Advances in Natural Language Processing, RANLP 2017, 2-8 September, 2017, Varna, Bulgaria.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014).
A convolutional neural network for modelling sentences.
CoRR, abs/1404.2188.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015).
Skip-thought vectors.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Le, Q. V. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
CoRR, abs/1405.4053.

Lin, D. (1998).
Automatic retrieval and clustering of similar words.
In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on*

Computational Linguistics - Volume 2, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.


Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.


Mitchell, J. and Lapata, M. (2010).
Composition in distributional models of semantics.
Cognitive Science, 34(8):1388–1439.


Partee, B., Meulen, A., and Wall, R. (1990).
Mathematical Methods in Linguistics.
Studies in Linguistics and Philosophy. Springer Netherlands.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011).
Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection.
In *Advances in Neural Information Processing Systems 24*.

van der Plas, L. and Tiedemann, J. (2006).
Finding synonyms using automatic word alignment and measures of distributional similarity.
In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016).
Towards universal paraphrastic sentence embeddings.
International Conference on Learning Representations, CoRR, abs/1511.08198.

Wu, H. and Zhou, M. (2003).

Optimizing synonym extraction using monolingual and bilingual resources.
In *In Proceedings of the second international workshop on Paraphrasing*, page 72.