DOI: 10.15439/2016F419
ACSIS, Vol. 8. ISSN 2300-5963

Word2vec Based System for Recognizing Partial Textual Entailment

Martin Víta
NLP Centre
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
Email: info@martinvita.eu

Vincent Kríž
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25, 118 00 Prague
Czech Republic
Email: kriz@ufal.mff.cuni.cz

***Abstract*—Recognizing textual entailment is typically considered as a binary decision task – whether a text T entails a hypothesis H. Thus, in case of a negative answer, it is not possible to express that H is "almost entailed" by T. Partial textual entailment provides one possible approach to this issue.**

This paper presents an attempt to use word2vec model for recognizing partial (faceted) textual entailment. The proposed approach does not rely on language dependent NLP tools and other linguistic resources, therefore it can be easily implemented in different language environments where word2vec models are available.

I. Introduction and Preliminaries

NOWADAYS, textual entailment belongs to intensively and deeply studied notions in NLP, with potentially many practical applications including paraphrase detection, multi-document summarization, machine translation evaluation, plagiarism detection, etc. In this section we provide a brief description of textual entailment, partial textual entailment, we mention word2vec model and present the main aim of this work.

A. Textual Entailment

Different definitions of textual entailment (abbr. as TE) and a systematic overview of this area can be found in an older but comprehensive paper [1]. *Recognizing textual entailment* (RTE for short) is a corresponding decision problem whether a given (coherent) text T entails a given text H (in this context often called a hypothesis). Currently, there exist several systems for RTE problem: an up-to-date list of them can be found at ACLwikiWeb[1]. Some of them were created in order to participate SemEval challenges.

Since RTE is a binary decision problem, in case of a negative result of RTE, i. e., when T does not entail H, it is not possible to state "how distant" is H from another hypothesis H', such that H' is entailed by T. From a different point of view, it is not possible to express that H is "almost entailed" by T in this setting. Partial textual entailment is one possible approach to this issue. The key elements of the idea of partial textual entailment were introduced in [2], although the notion of partial textual entailment was not explicitly mentioned in the paper. The motivation for partial textual entailment has naturally arised from the problem of (automatic) analysis of student responses in educational process.

B. Partial and Faceted Textual Entailment

According to [3], we say that an ordered pair $(T; H)$ forms a partial textual entailment (abbr. as PTE) if *a fragment* of the hypothesis H is entailed by T. In this definition, the notion of a fragment of the hypothesis is no more specified. Hence, the key question is how to decompose the hypothesis into fragments.

In [2], facets were introduced as special fragments: a facet is an ordered pair of words (f_1, f_2) that are contained in the hypothesis – accompanied by a semantic relation binding these words together. A simplified version of this approach – used in SemEval 2013 challenge – deals only with a pair of words *without* explicitly metioned semantic relation.

For example, if the hypothesis has the form of a sentence "The water was evaporated, leaving the salt.", one of corresponding facets is: (evaporated, water). Starting now, we are going to use only this simplified model.

The problem of recognizing faceted entailment can be stated as follows: *"Does the given text T express the same semantic relationship between the words f_1 and f_2 exhibited in H?"*

II. Note on a Related Work – Existing System for Faceted Entailment

Currently, there are only a very few systems for recognizing faceted entailment. In SemEval 2013 Task 7, only one system was submitted – a system of Levy et al. [3].

It consists of three components: *Exact Match, Lexical Inference, Syntactic Inference*. Exact match checks whether lemmas of words contained in the facet appear both in the text. The Lexical Inference is based on Resnik similarity [4] over WordNet [5]. The idea behind this module is to find out whether words semantically related (semantically similar) to those contained in the facet, occur also in then text.

The Syntactic Inference module is based on BIUTEE entailment engine that deals with dependence trees. The dependency

[1] http://aclweb.org/aclwiki

978-83-60810-90-3/$25.00©2016, IEEE

tree corresponding to a given facet is obtained from the dependency tree of the whole hypothesis using lowest common ancestor (LCA) of facet-nodes: it is just the path from one facet node to the second one via LCA node. This inference component has no paralel in our approach.

The best results of Levy et al. system were achieved in "Majority" configuration (Exact $\vee$ (Lexical $\wedge$ Syntactic)). In terms of F_1-measure, the scores vary from 0.765 to 0.816 depending on different scenarios.[2]

A. Main Aim of the Work

In this paper we present a novel system for recognizing partial/faceted textual entailment that is based on word2vec representations of the words contained in the text T and words contained in the facets.

The results of this monolingual setting can provide rough estimations of overall accuracy and other measures for intended cross-lingual modification that is briefly described in the last section of this text, thus this work can also be viewed as a prerequisite to cross-lingual faceted entailment.

B. Word2vec Model

Word2vec model belong to a class of distributed representations of words. The main attribute of distributed representations (proposed relatively long time ago, in the second half of 80th in [6]) is, that the representations of (semantically) similar words are close in the vector space.

Word2vec model arises from the idea of predicting the neighbours of a word using a neural network. There are two possible modes of predicting: distributed Skip-gram or Continuous Bag-of-Words (CBOW), see [7]. The CBOW idea is to predict the word "in the middle" from the surrounding words, whereas in Skip-gram model the training objective is to learn predicting its context in the same sentence. The (real number) vector representations of words correspond with the weights between input and first hidden layer in used deep feedforward network. The dimension of the target word2vec space is a parameter of the model.

III. Task Definition, Algorithm Description and Used Data

Recognizing faceted entailment is a binary classification task. The inputs are the text T and the hypothesis H along with the facet (f_1, f_2) of words contained in H. The output classes are *Expressed* and *Unaddressed*[3] (which means the semantic relationship between f_1 and f_2 is expressed explicitly or implicitly in T, or not, respectively).

Let us assume we have a word2vec model, i. e. for (almost) each word w we have its vector representation $r(w)$ in word2vec space of a given dimension, a text T an a facet (f_1, f_2). Parameters of our algorithm is a threshold α from the $(0, 1)$ interval.

A. Algorithm Description

The decision algorithm for faceted textual entailment (abbreviated as W2V in the following text) works in the following steps:

1) Split the text T into tokens $t_1, \ldots, t_n$.
2) Get the word2vec representations
$$r(t_1), \ldots, r(t_n), r(f_1), r(f_2)$$
whenever possible.
3) For f_1 select the word t_p such that $d(r(f_1), r(t_p))$ is equal to
$$\min\{d(r(f_1), r(t_k)) \mid 1 \leq k \leq n\},$$
where d is the standard cosine distance. For f_2 select analogously t_q. Roughly said, select two words in T that have the lowest distances to the facets in the sense of word2vec space.
4) If
$$\frac{d(r(f_1), r(t_p)) + d(r(f_2), r(t_q))}{2} \leq \alpha$$
than (f_1, f_2) is *Expressed* in T, otherwise (f_1, f_2) is *Unaddressed* by T. If some word of the facet is missing in the word2vec model, the result class is set to *Unaddressed*.

If the facet consists of more than two words (tokens), use this algorithm analogously for all of them.

The optimal value of α is obtained after experiments on training data – the selected value provides the best results of this algorithm in terms of overall accuracy. We will refer to this algorihtm as "W2V".

In addition, we will employ the trivial algorithm (that will be refered as "EXACTMATCH"): it returns *Expressed* in case that both words of the facet are contained in the text T, otherwise it returns *Unaddressed*. No lemmatization is taken into account since we are preparing a maximally language independent solution – in EXACTMATCH we deal only with word forms. This trival algorithm is used in order to treat with situations when a facet uses the same words as those contained in the text – but that are not contained in the word2vec model (for instance, correct words with a very low frequency).

B. Used Data and Word2vec Model

The evaluation was performed using a dataset derived from SciEntsBank corpus [8] that was used in the Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge at SemEval-2013 Task 7. This corpus is focused on previously mentioned domain of student response analysis. It contains scholar questions, reference answers and student responses. From the "practice point of view", the aim is to recognize whether the student's answer is at least partially correct. Transforming this issue to recognizing (partial) textual entailment environment models this situation: the role of the hypothesis H plays the reference answer and the role of the

[2] The comparison of our proposed system with this one was not provided due to missing information about the data used in each scenario.

[3] In the context of faceted entailment, "Expressed" and "Unaddressed" labels are used instead of "Entailed" and "Not entailed".

text T is played by the student's answer. If H is (partially) entailed by T, than student's answer is (partially) correct.

Let us illustrate it on the example.

QUESTION: *You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?*

STUDENT ANSWER: *Let the water evaporate and the salt is left behind.*

REFERENCE ANSWER: *The water was evaporated, leaving the salt.*

FACET: *(evaporated, water)*

As already mentioned, T is the student answer and the task is to decide whether the semantic relationship between "evaporated" and "water" is expressed in T. In this case, the relationship is "Expressed", thus the student answer can be regarded as partially correct.

In contrast, when student answers "I don't know." the facet *(evaporated, water)* is obviously not expressed.

The advantage of using this corpus is that facet extraction was already done and the faceted entailments were manually annotated. The SemEval-2013 Task 7 corpus is divided in two parts, training and test collections. The training collection contains 13145 pairs, the test collection contains 16263 pairs text-hypothesis (i. e. facets). As the texts T, we have considered just the student answers in all cases, no other texts (like parts of questions) were taken into the account.

While in case of "standard" recognizing textual entailment there are several training/test sets, for faceted/partial textual entailment, annotated corpora are very rare.

Word2vec model was built using the original implementation[4] over the TC Wikipedia[5]. Standard preprocessing issues were performed (e. g. lowercasing, punctuation removal). The model was obtained with the following basic parameters: the dimension was set to 200, the window was set to 5, the mode was CBOW.

IV. Results

Since recognizing faceted textual entailment is a binary classification task, the performance is measured in a standard way – obtaining precision, recall and F_1-measure scores over the test collection of SciEntsBank corpus. F_1-measure was chosen in order to compare our results with [3]. The threshold α in W2V algorithm was set to 0.555 – this value of the parameter maximizes the overall accuracy over the training collection.

The results are summarized in Table I, Table II and Table III. They were obtained by "official SemEval scripts"[6].

EXACTMATCH achieves relatively high precision at positive class, nevertheless it provides low recall – these characteristics correspond with "common sense" expectations. The combination of these two approaches leads to better results in F_1-measure than the W2V approach used separately.

TABLE I
W2V $\vee$ EXACTMATCH RESULTS

	Precision	**Recall**	**F_1-measure**
Expressed	0.661	0.811	0.729
Unaddressed	0.875	0.761	0.814

TABLE II
W2V RESULTS

	Precision	**Recall**	**F_1-measure**
Expressed	0.652	0.774	0.707
Unaddressed	0.854	0.761	0.805

V. Conclusion

We have presented a simple system for recognizing faceted textual entailment that is based on word embeddings: word2vec model in particular – other embeddings with similar characteristics (like GloVe) can be treatened in an analogous way.

Despite of its simplicity it provides reasonable results in terms of F_1-measure. The key features of this system are no need of other language resources in except of a relevant word2vec model and no usage of NLP tools. Thus it can be quickly implemented in any language where word2vec models can be created. The preparation of word2vec models requires only a collection of texts of a sufficient volume like Wikipedia in the corresponding language and/or a relevant web corpus (without any annotations).

Using word2vec model and our approach "simulates" the use of lemmatization in morphologically rich languages (since the cosine distance of a given word form and its lemma is usually very low – observed during experiments with Czech language), thus our approach would most likely achieve relatively comparable results also in other languages.

It can be straightforwardly implemented in different programming languages and environments – in our case, in `R` environment (enriched by `lsa` and `tm` packages) was used. Word2vec representations were stored in CSV format and were loaded into `R`.

Comparing to the mentioned approach of Levy et al. [3], our approach provides a comparable results in terms of overall accuracy – but it can be easily implemented also in "under-resourced" languages (where syntactic tools – as well as WordNet – are not available). Our proposed system approximately corresponds with the first two components of their system (Exact Match and Semantic Inference). The differences are summarized as follows: in [3], Exact Match contains lemmatization, in our approach lemmatization is not used.

TABLE III
EXACTMATCH RESULTS

	Precision	**Recall**	**F_1-measure**
Expressed	0.970	0.366	0.531
Unaddressed	0.731	0.993	0.842

[4] http://code.google.com/p/word2vec

[5] http://nlp.cs.nyu.edu/wikipedia-data/

[6] https://www.cs.york.ac.uk/semeval-2013/task7/data/uploads/datasets/semevaltask7code.zip

Semantic Inference module is in our setting "replaced" by low distances in word2vec space.

VI. Further Work

Our proposed system can serve as a baseline for further experiments.

Since word2vec models are able to capture many linguistic regularities [9], it is intended to employ rule based transformations on facet representations and subsequently determining whether transformed representations are contained in representation of T (for example, dealing with representations of hyper/hyponyms of words that forms the given facet, similarly as in [10]). These extension can be viewed as a paralel of Syntactic inference module of previously mentioned system.

In our presented approach, the threshold α stays constant in all cases and the distances $d(f_1, t_p)$ and $d(f_2, t_q)$ were simply combined into the mean, which was followingly tested against α. Other way of improving our system will be based on employing more features, e.g.:

- raw distances $d(f_1, t_p)$ and $d(f_2, t_q)$,
- the number of words in text between t_p and t_q,
- the angle between vectors $r(f_1) - r(f_2)$ and $r(t_p) - r(t_q)$,
- features obtained from hypernymy/hyponymy, synonymy and other attributes derived from WordNet data.
- ...

and using ML methods like SVM etc. We suppose that employing features (especially those arising from semantic resources like WordNet will help to improve both precision and recall).

Another part of further work is application-oriented: we are going to employ recognizig faceted entailment system in text summarization task (like that described in [11]) etc.

Note on the Cross-lingual Approach

As already mentioned, the proposed approach will be extended for using in a cross-lingual environment. It has been demonstrated in [12], paralel word2vec models can be used for of generating and extending dictionaries and phrase tables. The underlying idea is simple (with little assumptions about the languages involved): unknown word translations can be obtained by learning language structures over large monolingual data and mapping between languages on a small domain (in terms of the mapping).

More formally, let us have n word pairs and their vector representation $(x_i, z_i)_{i=1}^{n}$, where $x \in R^{d_1}$ is a vector representation of i-th word in the source language and $z \in R^{d_2}$ a vector representation of its translation. The goal is to find a matrix W such that Wx_i approximates z_i. The matrix W is obtained as a solution of an optimization problem:

$$\min_W \sum_{i=1}^{n} \|Wx_i - z_i\|^2.$$

In [12], this problem is solved with stochastic gradient descent.

At this moment, the modification of our algorithm for cross-lingual faceted entailment is straightforward: Having a facet (f_1, f_2) in the source language and the text T in the target language, then we take the vector representations (x_1, x_2) in source word2vec space, compute $(z_1, z_2) = (Wx_1, Wx_2)$ and determine representations of words that are the closest to z_1 and z_2 in the sense of cosine similarity in the target language word2vec model. The rest will be identical to the monolingual case.

Acknowledgment

The second author was supported by SVV - specific academic research - reg. no.: 260 333, which is funded by the Charles University in Prague.

References

[1] I. Androutsopoulos and P. Malakasiotis, "A survey of paraphrasing and textual entailment methods," *Journal of Artificial Intelligence Research*, pp. 135–187, 2010.

[2] R. D. Nielsen, W. Ward, and J. H. Martin, "Recognizing entailment in intelligent tutoring systems," *Natural Language Engineering*, vol. 15, no. 04, pp. 479–501, 2009.

[3] O. Levy, T. Zesch, I. Dagan, and I. Gurevych, "Recognizing partial textual entailment." in *ACL (2)*, 2013, pp. 451–455.

[4] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, 1995, pp. 448–453.

[5] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[6] D. R. G. H. R. Williams and G. Hinton, "Learning representations by back-propagating errors," *Nature*, pp. 523–533, 1986.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] M. O. Dzikovska, R. D. Nielsen, and C. Brew, "Towards effective tutorial feedback for explanation questions: A dataset and baselines," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 200–210.

[9] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations." in *HLT-NAACL*, 2013, pp. 746–751.

[10] J. A. Miñarro-Giménez, O. Marín-Alonso, and M. Samwald, "Applying deep learning techniques on medical corpora from the world wide web: a prototypical system and evaluation," *arXiv preprint arXiv:1502.03682*, 2015.

[11] K. Jassem and L. Pawluczuk, "Automatic summarization of polish news articles by sentence selection," in *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*, 2015, pp. 337–341. [Online]. Available: http://dx.doi.org/10.15439/2015F186

[12] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.