Distributional semantic models for detection of textual entailment

Yuri Bizzoni and Simon Dobnik

CLASP, University of Gothenburg, Sweden
{yuri.bizzoni,simon.dobnik}@gu.se

Abstract

We present our experiments on integrating and evaluating distributional semantics with the recognising textual entailment task (RTE). We consider entailment as semantic similarity between text and hypothesis coupled with additional heuristic, which can be either selecting the top scoring hypothesis or a pre-defined threshold. We show that a distributional model is particularly good at detecting entailment related to "world knowledge", and that aligning the hypothesis with the text improves detection of lexical entailment.

1. Introduction

Reasoning with natural language is one of the core tasks of computational linguistics, but it is an incredibly challenging one. This is because in natural language a valid inference or entailment can be made through several different relations and associated operations, some which follow from the formal linguistic structure (and can be captured through application of logical rules in theorem provers) and some that follow from the lexical properties of words. (Cooper et al., 1996)[1] provide a test-suite of examples of inference mostly of the first kind, which has been, for example, tested in (Sukkarieh, 2000), while the Recognizing Textual Entailment task (RTE) (Dagan et al., 2006) focuses on inference that is mostly of the second kind. An important difference between the two approaches is in their definition of entailment: while the approaches of the first kind only accept strict logic definition of entailment, RTE approach accepts a more relaxed definition, namely a hypothesis is a conclusion a human would most likely infer reading a text. The technologies proposed for this task can be further categorized according to the kind of inference they try to capture - for example, if they try to capture inferences that require external knowledge or not. Some systems narrow their scope only on entailments that can be completely deduced from the text and they rely on lexico-syntactic analysis without accessing a general knowledge base (see for example (Vanderwende et al., 2006). Systems trying to handle also external knowledge apply a variety of strategies, ranging from domain-specific frames that provide information about a given topic to techniques of dictionary mining to collect basic knowledge about entities through, for example, WordNet-like definitions and representations (Clark et al., 2007) that allow considering lexical relations among words. The shortcoming of such approaches is that the production of these resources is often expensive in terms of time and human work, and thus their availability is limited. NER taggers and algorithms for computing string or tree similarity are also used. In some cases (de Salvo Braz et al., 2005) transformation-based approaches have been proposed to transform a natural language expression into a formal equivalent, and making the latter more robust to mismatches. The approach can be seen as a constrain-based heuristic: fewer the necessary edits, higher the probability that a hypothesis is a valid inference from a text.

[1]Now available as an XML resource at http://www-nlp.stanford.edu/∼wcmac/downloads/fracas.xml.

2. Semantic vector spaces for RTE

Lexical or world knowledge information is an important aspect of making inference. For example, a system could fail to correctly label a text-hypothesis pair such as "John lives in Paris" – "John lives in France" for the simple fact that it doesn't know that Paris is in France. Therefore we expect that comparing a model of lexical meaning of the text and the hypothesis will be helpful in identifying entailment. Lexical meaning and semantic similarity of words in corpora are commonly modelled by distributional semantic vector spaces such as those built with `word2vec` (Mikolov et al., 2013). Therefore, an interesting question to ask is how successful distributional semantic models are in identifying entailment. Another question we have to address before, however, is how to integrate distributional lexical information into the inference task. Here we present two models and discuss and compare their results. We focus on the example mainly from the third PASCAL data-set for textual entailment (Marneffe et al., 2008).

2.1 Prerequisites

We used the gensim implementation of the Word2Vec model (Řehůřek and Sojka, 2010) with a pre-trained set of vectors computed on a subsection of Google News corpus of 100 billion words. This sect of vectors contains 300-dimension vectors for 3 million words and phrases. Word2Vec is a one-layered neural network widely used to create continuous distributional semantic spaces. A useful feature of such representations is that they allow creating mean vectors from the vectors of a group of words. Thus, if a Word2Vec model knows every word in a couple of documents, we can compute the distributional similarity of such two documents treating them as bags of words.

The simplest application of this property to RTE is thus to compute the mean vector of a text from the distributional vectors of its words and to draw its cosine similarity with the mean vector of words in its hypotheses.

The approach requires filtering out words the system has not seen before, but if the training dataset is large enough, as in our case, this is a relatively minor issue. As we will see, despite being so straightforward, the approach turns out useful for an RTE task.

2.2 Solution 1: text and hypothesis as a bag of words

Using this approach, text and hypotheses are reduced to a "point" in the semantic space, which is the mean vector of the vectors of their individual words (bag-of-words approach). The distributional profile of each word in a text evenly contributes to the distributional profile of the text. We studied the performance of our Word2Vec model on a set of a small set of cases (5) taken from the Stanford RTE3 dataset. For each text we provided at least two hypotheses, one which was an entailment and one which was not, as shown in *Example 1*:

T:	In 1956 Accardo won the Geneva Competition and in 1958 became the first prize winner of the Paganini Competition in Rome.	
H1:	Accardo won the Paganini Competition in Rome.	0.89
H2:	Accardo won the Nobel prize in literature.	0.78

For every case, we asked the model to label as entailment the hypothesis whose vector was most semantically similar to the text. Another heuristic is therefore required to separate the scored hypotheses. In the following examples we assume that the best ranked hypothesis is entailed. Here, the absolute similarity value of a text-hypothesis couple is thus not taken into account. However, other strategies are possible, for example n-best ranked hypotheses or an empirically pre-determined threshold.

Our model identified entailing hypothesis in all of the 5 test cases. We crafted these examples so that non-entailing hypotheses contained terms that are in our intuition semantically related but in terms of entailment distinct from the text ("Nobel prize" vs "Paganini Competition") which made the task relatively difficult. If the difference between the text and the wrong hypothesis is essentially lexical, distributionality gives good results: in the example above, the similarity between the text and H1 is 0.89, while the similarity between the text and H2 is 0.78.

It is important to note that using a semantic space for RTE is not just a complicated variant of a string similarity measure. It allows to handle various degrees of semantic relatedness between words. For example a text - hypothesis couple like "A poet won the Nobel prize" and "An artist was awarded with a prize" holds a relatively high degree of similarity because "artist – poet", "won" – "awarded" and "Nobel" – "prize" are near in the semantic space. Hence, the strategy of looking for the most similar hypothesis can work even when no common terms are shared between text and entailment.

The following examples demonstrate that semantic spaces successfully contribute in modelling world knowledge for entailment. Example 2:

T:	In 1956 Accardo won the Geneva Competition and in 1958 became the first prize winner of the Paganini in Rome.	
H1:	Accardo won the Paganini in an Italian city.	0.81
H2:	Accardo won the Paganini in a Chinese city.	0.79

The first hypothesis' vector is closer to the text than the second's: "Italian" – "Rome" vs "Chinese". Example 3:

T:	She was transferred again to Navy when the American Civil War began, in 1861.	
H1:	The American Civil War started in 1861	0.82
H2:	The American Civil War started in the XIX century	0.8
H3:	The American Civil conflict started in 1861	0.79
H4:	The African Civil War started in 1861	0.78
H4:	The American Civil War started in XX century	0.76

This similarity measure might be misleading. For example, the text "Robinson was born in Ireland" and its entailed hypothesis "Robinson was born on planet Earth" gets a similarity score of only 0.54, because neither "planet" nor "Earth" are very similar to "Ireland" in our semantic space. Reversely, the text "All domestic animals eat plants that have scientifically proven medicinal properties" and the hypothesis "All wild mountain animals eat plants that have scientifically proven medicinal properties" have a high distributional similarity (0.94), but they don't entail each other. The approach also does not take into account the sequence of words/syntax: "Accardo won the Paganini Competition in Rome" and "Rome won the Paganini Competition in Accardo" are equally good hypotheses for the text in the first example.

2.3 Solution 2: taking into account word order

Several approaches in RTE introduce alignment. The intuition behind this is that the hypothesis that most closely matches the order of information in the text is the preferred one. In our experiment, we used a Python implementation of the Needlmenann-Wunsch alignment algorithm (Needleman and Wunsch, 1970), previously used in (Bizzoni, 2015; Bizzoni and Reboul, 2016), to which we integrated a similarity function based on Word2Vec. While alignment methods have been applied in RTE tasks before, this is the first time, to the best of our knowledge, that this approach is combined with a distributional semantic space to detect entailment.

The alignment algorithm tries to align an hypothesis to the text - the more text and hypothesis differ, the more difficult alignment will be. In this case, the similarity measure that the Needlemann-Wunsch will use to align two sentences word by word is distributional similarity. Each word in the text is compared to each word in the hypothesis and their cosine similarity is virtually stored into a two dimensional matrix. From this information, the algorithm computes the optimal alignment between the two sequences as the least expensive path through the matrix. Every time the path through the matrix is not diagonal, for example if an element in the two sequences doesn't find a match into the other sequence, the alignment score of the following two elements decreases, so the optimal path minimizes the number of the un-matched elements. Once the alignment is performed, we have a set of pairs of words from text and hypothesis that are aligned, and a set of words that did not find a match and are therefore not aligned. We then compute the entailment score of a hypothesis by summing the cosine similarity of every aligned couple and by dividing the sum by the length of the hypothesis.

Using this method, words with similar distribution will tend to be aligned, if an excessive variation in the structure

of the two sentences does not prohibit such alignment. For example, given the text "The Japanese surrendered on May 25, 1945" and the hypothesis "An Asian country surrendered in Spring 1945", the algorithm performs the following alignment.

Text	Hypothesis	Align score	Sim score
The	An	0.58	0.58
Japanese	Asian	0.06	0.53
	country	0.05	0
surrendered	surrendered	0.03	1.
on	in	0.25	0.39
May	Spring	0.11	0.3
25		0.0	0
1945	1945	0.12	1

The third column gives the alignment scores and the fourth the distributional similarity between aligned words. Alignment scores can be hard to follow without the Needlemann-Wunsch alignment grid, but it can be seen how the first alignment mirrors the basic similarity score, while subsequent alignments tend to become weaker around missing matches (such as around "country" and "25") and rise again at points that are central to a series of aligned elements (such as the "on" - "in" alignment). World-knowledge benefits are still preserved as "Japanese" aligns with "Asian" and "May" with "Spring".

In this model, we chose threshold as our heuristic for selecting the entailed hypothesis. Since the scores are based on cosine similarity, a threshold 0.5 indicates a midway between a perfect semantic similarity and a complete dissimilarity. Hence, a hypothesis that scores higher than 0.5 is marked as entailment. In the previous example the sum of the Similarity scores is 3.81 which, divided by 7 (the number of words in the hypothesis) returns 0.54: a score we deem as entailing. Table 1 shows some further text-hypothesis pairs to which this method was applied and their entailment scores.

The alignment method also has shortcomings: verbose or redundant hypotheses can lower the overall score and entailments formulated in particular ways can present difficulties for alignment, including the previously mentioned active – passive pairs.

3. Disambiguating figurative language

An interesting subset of RTE cases are the text-hypothesis pairs containing figurative language. It is intriguing to inspect to what extent this system is able to align, for example, a figurative hypothesis with the elements in the text that allow its interpretation. Let's consider the following text: "In the new version of the game Zorgs is the main character and he is very clever" and its figurative hypothesis: "Zorgs is a fox". The alignment of the last part of the text with the figurative hypothesis is as follows:

Zorgs is the main character and he is very clever
Zorgs is a fox

The algorithm does not align "is a fox" with "is the main character". Instead, it detects a relation between cleverness and foxes and correctly aligns "is a fox" with "is very clever". The overall entailment score is 0.5. But what happens if we produce a metaphorical hypothesis that does not seem to fit anywhere in the text? Let's consider the hypothesis "Zorgs is an old hamburger". Here, the relation with "clever" is lost: "old hamburger" is aligned with "he is" and the score sinks to 0.4.

Here is another alignment example:

He asked Jim for help but Jim had a heart of stone
Jim was cruel and indifferent

"heart of stone" aligns with "cruel and indifferent". Again, if we change the text to "He asked Jim for help and Jim had a heart of gold", the association stone–indifference is lost and the system tentatively aligns "cruel and indifferent" with "help and Jim had", lowering the overall score. The "correct" alignment is restored if we change the hypothesis to "Jim was a generous person", and so on.

Given a set of attributes in text and a metaphor in hypothesis, the aligner helps us to identify which attributes are consistent with the metaphor. In the example with the text "she was busy and smart while he was quick and had a huge memory" and the hypothesis "she was a bee he was an elephant", "bee" aligns with "busy" and "an elephant" aligns with "huge memory" (for scores see Table 1). In several of these examples, the score is low or under the threshold of entailment, which is in a way reassuring: metaphors are, after all, semantic inconsistencies.

4. Evaluation

We ran some small scale evaluation of our system on two datasets of interest, the whole RTE dataset edited by Stanford we used to pick the previous examples (482 examples in total) and the FraCas test set (236 examples in total) (Cooper et al., 1996). Both test sets divide examples 'entailment', 'contradiction' and 'unknown'. FraCas has also 'undefined'. These are problematic case where it could be yes or no if additional information is added to it. In a first experiment, we filtered out the 'unknown' and 'undefined' cases, running our model on entailments and contradictions. We achieved a precision of 82% with FraCas and 74% with Stanford RTE dataset.

It is of interest to note, though, that the model performs worse if we add examples marked as unknown than when dealing only with contradictions. In a second experiment we chose to label both 'unknowns' and 'contradictions' as non-entailments: in this case, precision scales down to little more than 59% in both datasets.

This outcome might be due to the fact that contradictions tend to contain elements that complicate the construction of the final alignment and thus lower the overall similarity score, while unknowns mostly play on a lack of information, as was shortly discussed in the paper. We can find nonetheless a polarity in the average score our model assigns to entailments and 'unknowns': entailments have an average score of 0.68 and 'unknowns' of 0.58.

5. Conclusion and future work

We presented two methods of identifying entailment between sentences based on cosine similarity in distributional vector spaces. In the first method we take a bag of words approach, while in the second method we add the notion of word order in the form of alignment between text and hypothesis which allows us to successfully identify a wider

In 1956 Accardo won the Geneva Competition and in 1958 became the first prize winner of the Paganini Competition in Rome.		
Accardo won the Paganini in Rome.	0.9	Entailed
Accardo won the Paganini in an Italian city.	0.6	Entailed
Accardo won the Paganini in a city.	0.6	Entailed
Accardo won the Paganini in a Chinese city.	0.5	Non-Entailed?
An Italian city won the Paganini in Accardo.	0.4	Non-Entailed
In the Super Nintendo Entertainment System release of the game Final Fantasy III, Biggs' name was Vicks.		
The Super Nintendo Entertainment System released the game Final Fantasy III.	0.81	Entailed
The game Final Fantasy III released the Super Nintendo Entertainment System.	0.41	Not-entailed
KOOG was Utah's first Pax affiliate, and changed its call letters to KUPX in February 1998.		
KUPX was an affiliate of KOOG.	0.3	Not-entailed
The Gasp Peninsula or just the Gasp (la Gaspsie in French) is a North American peninsula on the south shore of the Saint Lawrence River, in Quebec.		
The Gasp Peninsula is located in Quebec.	0.9	Entailed
Edison , Dickson and the other employees of Edison's laboratory made progress on the design to a point.		
Dickson worked for Edison.	0.6	Entailed
Buckley's Mixture is a cough syrup invented in 1919 (and still produced today) noted for its extremely bitter taste.		
Buckley's Mixture is a remedy against cough.	0.77	Entailed
She was busy and smart while he was quick and had a huge memory.		
She was a bee and he was an elephant.	0.6	Entailed

Table 1: Text-hypothesis pairs evaluated using Solution 2 with alignment

number of cases related to word ordering. We also demonstrate how using distributional semantics and alignment can be very interesting in particular text-hypothesis cases, those containing figurative language.

The results of the proposals in predicting entailment that is related to lexical similarity between sentences and the order in which information is communicated in them are very encouraging on the examples investigated. As stated at the beginning, detecting textual entailment is a multi-faceted challenge and the approach based on lexical similarity would do less well on examples of entailment that are based on the structure of logical forms. However, combined with logical tools it may lead to positive results. Recently, there has been significant advances in approaches that allow compositionality of vector spaces (Clark, 2015). Integration of such approaches with our proposals could enhance their performance.

So far the system has been evaluated on several hand-picked and modified examples from the 3rd PASCAL dataset for textual entailment. In our immediate forthcoming work we will do a systematic evaluation of the systems on the RTE suite which will give us a more complete picture of its performance and identify areas of good and bad performance.

References

Yuri Bizzoni and Marianne Reboul. 2016. Projet Odysseus, outil d'études comparatives du traductologue. In *Digital Humanities (DH2016)*.

Yuri Bizzoni. 2015. The Italian Homer. Master's thesis, University of Pisa.

Peter Clark, William R. Murray, John Thompson, Phil Harrison, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in RTE3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 54–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics — second edition*, chapter 16, pages 493–522. Wiley – Blackwell.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium. ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Rodrigo de Salvo Braz, Roxana Girju, Vasin Punyakanok, Dan Roth, and Mark Sammons. 2005. An inference model for semantic entailment in natural language. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *MLCW*,

volume 3944 of *Lecture Notes in Computer Science*, pages 261–286. Springer.

Marie-Catherine De Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *In ACL 2008*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

S. Needleman and C. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Jana Zuheir Sukkarieh. 2000. *Natural language for knowledge representation*. Ph.D. thesis, Computer Laboratory, Churchill College, University of Cambridge, Cambridge, UK, April.

Lucy Vanderwende, Arul Menezes, and Rion Snow. 2006. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.