Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Semi-Automatic Construction of a Textual Entailment Dataset: Selecting Candidates with Vector Space Models

Erick R. Fonseca[1], **Sandra M. Aluísio**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brazil

{erickrf,sandra}@icmc.usp.br

Abstract. *Recognizing Textual Entailment (RTE) is an NLP task aimed at detecting whether the meaning of a given piece of text entails the meaning of another one. Despite its relevance to many NLP areas, it has been scarcely explored in Portuguese, mainly due to the lack of labeled data. A dataset for RTE must contain both positive and negative examples of entailment, and neither should be obvious: negative examples shouldn't be completely unrelated texts and positive examples shouldn't be too similar. We report here an ongoing work to address this difficulty using Vector Space Models (VSMs) to select candidate pairs from news clusters. We compare three different VSMs, and show that Latent Dirichlet Allocation achieves promising results, yielding both good positive and negative examples.*

1. Introduction

Recognizing Textual Entailment (RTE) is a Natural Language Processing (NLP) task aimed at determining when a given piece of text T entails the meaning of a hypothesis H. It is useful in many NLP applications, such as Question Answering, Automatic Summarization, Information Extraction and others [Androutsopoulos and Malakasiotis 2010, Dagan et al. 2013].

For example, (1) entails the meaning of both (2) and (3). The last two sentences entail each other, forming a *paraphrase* relationship. Paraphrases can be seen as a special case of entailment, occurring when both pieces of text have essentially the same content.

(1) For the accession of new contracts, the closing date was kept on the 30th, as previously informed.

(2) The deadline for new contracts accession is on the 30th.

(3) New contracts can be accessed until the 30th.

The exact definition of entailment in the NLP research community is rather subjective. The widely accepted notion is that T entails H when a person reading T would affirm that H is most likely true [Dagan et al. 2009]. We also follow this view here.

In order to automatically recognize when T entails H (or refute this possibility), the NLP community has come up with many different strategies, with no single one emerging as the best [Dagan et al. 2013]. Virtually all of them, however, need labeled data in order to calibrate system parameters; moreover, a labeled dataset is necessary to evaluate systems' performance in a standardized benchmark.

The lack of labeled RTE data is a major obstacle to research in this area for Portuguese. Obtaining an entailment dataset, however, is not a simple task: while in some other areas, such as part-of-speech tagging and named entity recognition, it suffices to pick texts and have a group of annotators tag them, some other points must be taken into account when it comes to RTE.

First, each item considered in RTE is actually a pair of text passages. Such pairs will hardly be found in a single text document; instead, they must be collected from different but related sources. Some possibilities are news articles grouped by subject or different translations of the same original text; since each one in these cases is a description of the same event, a sentence in one text may entail a sentence in another one.

Second, the presence or absence of an entailment relation must not be obvious. In other words, T and H should be somewhat similar to each other, especially when there is no entailment; conversely, if T entails H, they should bear some differences. Thus, while (4) is entailed by (1), it would be a bad example in an RTE dataset, not reflecting the difficulty of the task, since the only change was the replacement of a word by a synonym.

(4) For the accession of new contracts, the closing date was maintained on the 30th, as previously informed.

If this point is kept in mind during the construction of an RTE dataset, it will be of greater practical use. For example, consider the case of a Question Answering system which has formulated a candidate answer to a question, and needs to check whether it is entailed by a text from some trusted source. Even if the text does entail the answer, it will probably have different words and a different syntactic structure.

In this work, we aimed at obtaining nontrivial RTE pairs in Portuguese in order to create a gold standard dataset. We took advantage of the news clusters provided by the Google News service[1], which groups news by subject, making them ideal for extracting RTE candidate pairs. We then used vector space models [Turney and Pantel 2010] to select similar sentences from different documents.

A full manual revision of the extracted pairs still needs to be carried out in order to label them according to the relation they display (entailment, paraphrase or neither one) and filter out bad candidates (that is, pairs where T and H are either too similar or too different). A preliminary analysis of a sample, however, showed that this method yields very promising data, containing good positive and negative examples.

The remainder of this paper is organized as follows. Section 2 discusses relevant related work and gaps in the area. Section 3 describes our method. Our results are presented and discussed in Section 4, and Section 5 shows our final conclusions.

2. Related Work

In the first PASCAL RTE Challenge[2] [Dagan et al. 2005], the main event dedicated to RTE research, a dataset was created by composing subsets related to different NLP applications, such as Information Retrieval or Automatic Summarization. Each subset was collected differently, sometimes with significant human labor. While in subsequent edi-

[1] https://news.google.com/

[2] http://pascallin2.ecs.soton.ac.uk/Challenges/

tions the organizers improved the dataset generation process [Dagan et al. 2009], some limitations still existed.

For example, in the Information Retrieval setting, annotators formulated queries to search engines. Each query was treated as a hypothesis and a candidate text was selected from one of the returned documents to form a (T, H) pair. More related to our process, in the Automatic Summarization task, annotators examined the summary of a news cluster and selected sentences (outside the summary) with high lexical overlap with it. Besides the labor required from annotators, this process also suffers from a possible bias in the way humans select pairs.

[Dolan et al. 2004] present the Microsoft Research Paraphrase (MSRP) Corpus, the *de facto* standard dataset for training and evaluating paraphrase detection systems. The paraphrase pairs were collected using two strategies, both exploring clusters of related news. The first one selected sentence pairs from the same cluster according to their edit distance. The second one compared the first sentence in each news article (which often summarizes the article's content) and selected the ones with some lexical overlap. Some filtering criteria discarded pairs with very different lengths. The dataset was revised by human annotators afterward.

In order to bootstrap an RTE dataset, [Hickl et al. 2006] took advantage of the fact that the headline of a news article is usually a reduced version of its first sentence. Their method then treats the first sentence of an article as T and its headline as H, and assumes it is a positive pair. For negative examples, they used two methods: the first selects two consecutive sentences mentioning the same named entity, and the second one selects any pair of consecutive sentences linked by contrastive connectives such as *even though* or *although*. A sample of the data was analyzed manually, and over 90% of the pairs had the expected class (positive or negative). However, their bootstrapped dataset suffers from the problems mentioned earlier: positive pairs are too similar while negative ones are very different.

The SICK dataset [Marelli et al. 2014] was created with a different approach. As a first step, sentence pairs describing the same image or video fragment were collected. Then, altered versions of these sentences were generated (with changes such as a negation or a noun replacement) and added to a pool. The pairs in the final dataset were picked by combining either sentences that originally described the same picture/video or not, and both could be the original or altered versions, which allowed a great variability in the dataset. Each pair was annotated by several reviewers using a crowdsourcing platform.

Concerning Portuguese, there is the AVE[3] (Answer Validation Exercise) dataset. It consisted in evaluating whether answers returned by QA systems [Rodrigo et al. 2009] were entailed by a supporting text, returned together with the answer itself. Thus, the supporting text was interpreted as T and the full sentence that answers the question as H. For example, for a question like "*What is the capital of Croatia?*", and a simple answer such as "*Zagreb*", H would be "*The capital of Croatia is Zagreb*".

While this reflects a real world application, it is limited to the QA scenario and to the shortcomings of the participating systems. Most problems can be seen in negative examples: in some cases, T is completely unrelated to H or can't be understood out of

[3] `http://nlp.uned.es/clef-qa/repository/ave.php`

context; and in the case of some wrong answers, H is either agrammatical or doesn't even make sense. Consider, for example, the pair (5) and (6), which was created based on the answer to the question "*A que se refere o termo "Les Six" em música?*" (What does the term "Les Six" refers to in music?): H doesn't make any sense and is completely unrelated to T.

(5) **[T]** Não gosta de falar de carreira, porque diz que é um termo com que não se identifica.
He doesn't like to talk about career, because he says it is a term he doesn't identify with.

(6) **[H]** "Les Six" são que é.
"Les six" are what it is.

3. Methodology

In order to extract RTE pairs, we examined clusters from Google News. Exploiting news clusters for paraphrase or entailment pairs acquisition is not a novel idea, having already been successfully explored before [Dolan et al. 2004, Dagan et al. 2005, Barzilay and Lee 2003]. However, instead of having human annotators select pairs based on word overlap, we employ Vector Space Models to suggest candidates.

3.1. Vector Space Models

Vector Space Models (VSMs) refer to a family of methods that map words or documents to a multidimensional space [Turney and Pantel 2010], such that each word or document is associated with a numeric vector. VSMs have a long history in NLP, being especially useful in the field of Information Retrieval [Manning et al. 2008].

The main purpose of VSMs is to model similarity: similar documents are mapped to similar vectors. The similarity between two vectors is usually expressed as their cosine; similarity between documents can be understood as how much common content they have. Robust methods should be able to measure it even when two documents don't have many words in common.

For example, suppose one document has many mentions of *money* and *investment*, while another one doesn't have these words, but has occurrences of *dollar* and *economy*. If a VSM has seen enough examples, it should indicate that these documents probably have a substantial degree of similarity. Thus, one advantage of using a VSM is its capability of identifying similar documents without relying too much on lexical overlap.

VSMs are generated after analyzing large corpora, without the need of any kind of annotation. In our experiments, we tried three different VSM methods: Latent Semantic Indexing (LSI) [Landauer and Dumais 1997], Latent Dirichlet Allocation (LDA) [Blei et al. 2003] and Random Projections (RP) [Sahlgren 2005]. These three models are based on computing word frequencies across documents, and applying linear algebra techniques to a matrix of word counts. After a VSM has been generated, it can project new documents into vectors, based on the words that occur in them.

Figure 1 illustrates how the similarity between two sentences (each viewed as a document) is calculated. Each document is given as input to the VSM, resulting in a real valued feature vector. Their similarity is calculated as the cosine of their vectors.

In recent years, new kinds of VSMs based on neural networks have emerged in NLP [Pennington et al. 2014, Mikolov et al. 2013]. These models, however, tend to focus

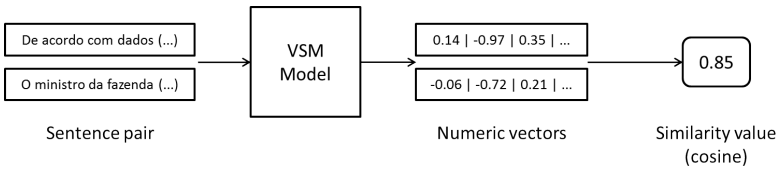

Figure 1. Obtaining a similarity value for two sentences

on the representation of words (also called *embeddings*) rather than documents. While methods for combining them into sentences or larger texts do exist [Le and Mikolov 2014, Socher et al. 2013], here we chose to use models based on occurrence counting, since they are simpler and more commonly used in retrieval-like tasks.

3.2. Candidate Pair Extraction

We generated the VSMs from a corpus of 8 months of news articles (from February to mid-October 2014, approximately 220 thousand articles and 100 million tokens) collected from the G1 website[4]. Since we wanted to work with sentences and not whole texts as RTE candidate pairs[5], we split the articles' texts into sentences and treated each one as a document from the VSM point of view (resulting in around 3.6 million documents). Although RTE candidates were not extracted from this corpus, using sentences as our document-level unit allows the VSMs to perform a more fine grained analysis.

Articles were preprocessed using common procedures: we converted the whole texts to lower case, removed stopwords, words occurring in less than 5 sentences or more than 50% of them. All three models were generated with 100 dimensions (or topics). For simplicity, we did not extract n-grams from the text, although this is worth investigating in future experiments.

After the models were generated, we used them to pick RTE candidates from our Google News corpus. This second corpus is composed of 329 clusters, each one containing on average 17.6 texts, totaling 90,310 sentences and around 2.4 million tokens. All texts were split into sentences, and each sentence was projected into the VSM and compared with others from the same cluster (in decreasing order of cosine similarity) until an appropriate pair was found.

A pair was considered appropriate when its cosine similarity was above a minimum threshold s_{min} and below a maximum s_{max}. Additionally, at least some proportion α of the words appearing in each sentence should not appear in the other. These conditions should select pairs that bear some similarities, but not too much. In order to avoid a bias towards a given topic, we limited the extraction process so that it could only pick two pairs from each cluster.

Figure 2 illustrates our whole procedure: first, the VSM is generated from a large corpus. Then, each news cluster is examined using the VSM in order to identify similar sentences. The resulting pairs are candidates to our RTE dataset.

[4] `http://g1.globo.com/`

[5] RTE candidate pairs do not need to be sentences. In fact, in the RTE Challenge datasets, some pairs had whole paragraphs as the T component. However, we want to explore here the simpler case of both T

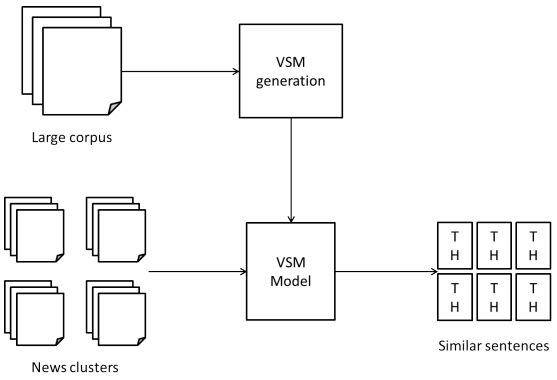

Figure 2. Similar sentence extraction process.

Ideally, we would like to investigate the effect of using the three different VSMs in tandem with different values for s_{min}, s_{max} and α. However, the dataset evaluation is costly, since it must be performed by human judges. Therefore, we only evaluated one configuration for all three models. After observing around 50 sentence pairs with varying parameter values, we chose to use $s_{min} = 0.65$, $s_{max} = 0.9$ and $\alpha = 0.35$. Lower values of s_{min} mostly allowed too different pairs; higher values of s_{max} allowed too similar ones. The parameter α only has significant impact on higher similarity values, when it filters out too similar candidates.

4. Results and Discussion

After generating datasets with each of the VSM methods, a human analysis was necessary in order to evaluate their quality and to ascertain whether there was or not an entailment relation in each pair.

So far, only a pilot analysis has been performed by a single judge. Its aim was not simply to annotate the data, but also to familiarize the annotator with the kind of relations that may be found in the pairs. This is essential in order to elaborate an annotation manual which minimizes the subjectivity of the task. In a later stage, we plan to have a group of annotators analyze all pairs, effectively creating the gold standard. Each pair was assigned one of six classes:

Entailment One sentence entails the other. Since the VSM similarities are symmetrical, the annotator had to indicate the direction of the entailment (whether the first sentence entailed the second or vice-versa).

Paraphrase Both sentences have essentially the same meaning.

No relation Sentences have some similarity, but neither entails the other. We call this case a negative example for the RTE dataset.

Large overlap, but no relation This happens when the two sentences share most of their content, but each one has some information that the other doesn't.

and H being sentences.

Too similar Sentences are too similar in syntactic structure and wording to be useful as an RTE example. Still, they usually display a paraphrase relationship.

Too different Sentences are too different to be useful as an RTE example.

The first three classes are interesting as RTE examples and should be kept in the final dataset, while the last two should be discarded. The notion of when a pair is too similar or too different is rather subjective, but it is important to filter out such cases. An example of a too different pair is shown in (7) and (8); (9) and (10) show a too similar one.

(7) Senado adia votação do novo indexador de dívidas dos Estados.
Senate postpones voting for the new debt index of the states.

(8) O objetivo é fixar em lei as regras para que os estados usem os recursos dos depósitos judiciais.
The objective is to establish in law the rules for the states to use resources from court deposits.

(9) A segunda partida acontece no próximo domingo, dia 03, na Arena Joinville.
The return leg takes place next Sunday, the 3rd, in Arena Joinville.

(10) A partida de volta em Joinville acontece no próximo domingo (3), às 16h.
The return leg in Joinville takes place next Sunday (the 3rd), at 16h.

The fourth class, however, requires more careful investigation. One option is to follow a stricter view and regard pairs in this category as not having an entailment relation; another one would be a more permissive interpretation that would consider that T entails H even if H has some piece of information not found in T.

An example of overlapping sentences is the pair (11) and (12). The first one informs that the game took place on a Wednesday, and that the winning team was Internacional. The second one doesn't mention the team's name, but tells that the game was on its home.

(11) O Internacional manteve a boa fase e venceu o Strongest por 1 a 0 nesta quarta-feira, garantindo a liderança do Grupo 4 da Libertadores.
The Internacional [6] *kept the momentum and won the Strongest 1-0 this Wednesday, guaranteeing the leadership of the Group 4 of Libertadores.*

(12) Em casa, a equipe gaúcha derrotou o The Strongest, por 1 a 0, e garantiu a primeira colocação do Grupo 4 da Copa Libertadores.
Playing at home, the gaúcho[7] *team defeated The Strongest 1-0 and guaranteed the first place in the Group 4 of the Libertadores Cup.*

While in some situations the missing information might make all the difference (e.g., in a query about the day a game was played), we believe that the pros and cons of treating such pairs as positive or negative should be carefully analyzed.

The results of the manual analysis over a sample of 100 pairs produced by each VSM are summarized in Table 1. They suggest that RP tends to select pairs with higher

[6] Soccer team
[7] From Rio Grande do Sul state in Brazil.

similarity: it has the most cases of pairs discarded for being too similar and the most paraphrases, and was the only method to select more positive entailment cases than negative ones.

LSI, on the other hand, has the highest number of "too different" pairs and the lowest number of overlaps. LDA appears to have a good balance between both extremes, with the lowest total amount of pairs that need to be discarded either for excessive similarity or difference.

Class	LDA	LSI	RP
Entailment	12	15	16
Paraphrase	5	3	10
No relation	35	30	14
Overlap	20	9	17
Too similar	4	0	12
Too different	24	43	31

Table 1. Counts of each class in the data generated by the VSMs

The number of negative examples extracted by LDA is a very promising result, since our main concern was the lack of methods for extracting good quality negative RTE pairs. We want to reinforce that a *negative pair* here means not only the absence of an entailment relation, but also that T and H talk about related subjects.

All methods selected a low number of positive entailment cases[8]. While we expected higher figures, there are still ways to circumvent this problem. One way would be to extract pairs from some news clusters using LDA, thus yielding more negative examples, and use Random Projections on others, which would give us more positive pairs. Also, since our final goal is obtaining an RTE dataset in Portuguese, not bound to a specific method, we could experiment with some strategies found in the literature for obtaining positive entailment pairs.

5. Conclusions

In this work, we have presented a strategy for obtaining candidate pairs to build a gold standard RTE dataset for Portuguese. We focused on avoiding trivial pairs, such as sentences too similar or too unrelated to each other. For that, we used three different Vector Space Models to select similar sentences from news clusters.

A preliminary analysis of a sample of the data showed that the proposed method can achieve good results, especially with LDA. Particularly concerning negative cases, which are scarcely explored in the literature, it can pick many good quality examples.

As future work in this line of research, we plan to carry out the manual annotation of the whole dataset in order to provide the Portuguese NLP community with a reliable RTE dataset. This dataset could then be used as a benchmark to train and evaluate systems.

The code used for the experiments reported here as well as the annotated data can be found at `http://nilc.icmc.usp.br/rte-bootstrapper/`.

[8]Note that, for RTE purposes, paraphrases can also be considered positive entailment pairs.

References

[Androutsopoulos and Malakasiotis 2010] Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

[Barzilay and Lee 2003] Barzilay, R. and Lee, L. (2003). Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23.

[Blei et al. 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

[Dagan et al. 2009] Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

[Dagan et al. 2005] Dagan, I., Glickman, O., Gan, R., and Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL challenges on Recognizing Textual Entailment*.

[Dagan et al. 2013] Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

[Dolan et al. 2004] Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356.

[Hickl et al. 2006] Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B., and Shi, Y. (2006). Recognizing Textual Entailment with LCC's GROUNDHOG System. In *Proceedings of the Second PASCAL challenges on Recognizing Textual Entailment*, pages 80–85.

[Landauer and Dumais 1997] Landauer, T. K. and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

[Le and Mikolov 2014] Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31 st International Conference on Machine Learning*.

[Manning et al. 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Marelli et al. 2014] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

[Mikolov et al. 2013] Mikolov, T., tau Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.

[Pennington et al. 2014] Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[Rodrigo et al. 2009] Rodrigo, A., Peñas, A., and Verdejo, F. (2009). Overview of the answer validation exercise 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *Lecture Notes in Computer Science*, pages 296–313. Springer Berlin Heidelberg.

[Sahlgren 2005] Sahlgren, M. (2005). An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

[Socher et al. 2013] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

[Turney and Pantel 2010] Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.