Unsupervised and Language-Independent Method to Recognize Textual Entailment by Generality

Sebastião Pais
MINES ParisTech
Centre de Recherche en Informatique
77305 Fontainebleau, France
pais@cri.ensmp.fr

Gaël Dias and **Rumen Moraliyski**
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
firstname.lastname@unicaen.fr

João Cordeiro
University of Beira Interior
HULTIG
6200 Covilhã, Portugal
jpaulo@di.ubi.pt

Abstract

In this work we introduce a particular case of textual entailment (TE), namely *Textual Entailment by Generality* (TEG). In text, there are different kinds of entailment yielded from different types of implicative reasoning (lexical, syntactic, common sense based), but here we focus just on TEG, which can be defined as an entailment from a specific statement towards a relatively more general one. Therefore, we have $T \overset{G}{\rightarrow} H$ whenever the premise T entails the hypothesis H, the hypothesis being more general than the premise. We propose an unsupervised and language-independent method to recognize TEGs, given a pair $\langle T, H \rangle$ in an entailment relation. We have evaluated our proposal through two experiments: (a) Test on $T \overset{G}{\rightarrow} H$ English pairs, where we know that TEG holds; (b) Test on $T \rightarrow H$ Portuguese pairs, randomly selected with 60% of TEGs and 40% of TE without generality dependency (TEnG).

1. Introduction

TE aims to capture major semantic inference needs across applications in Natural Language Processing (NLP). Automatic identification of TEs has become a relevant issue promoted by the series of challenges on Recognizing Textual Entailment (RTE), where it is defined as a directional relationship between pairs of text expressions denoted by T (the entailing *"Text"*) and H (the entailed *"Hypothesis"*). We say that T entails H if humans reading T would typically infer that H is most likely true (Dagan et al., 2005). Basically, RTE is the task of deciding, given two text fragments, whether the meaning of one of the texts is entailed (can be inferred) from the other one. As noted by Dagan et al. (2005), this definition is based on common human understanding of language, much like the definition of any other language understanding task. Accordingly, it enables the creation of gold-standard evaluation data sets for the task, where humans can judge whether the entailment relation holds for a given $\langle T, H \rangle$ pair. This setting is analogous to the creation of gold standards for other text understanding applications like Question Answering (QA) and Information Extraction (IE), where human annotators are asked to judge whether the target answer or relation can indeed be inferred from a candidate text.

We introduce the TEG paradigm, which can be defined as the entailment from a specific sentence towards a more general one. For example, the pair $\langle S_1, S_2 \rangle$, taken from the RTE-1 corpus, naturally evidences that S_1 entails/implies S_2, and the latter is more general. Therefore, we have TEG from S_1 to S_2, denoted as: $S_1 \overset{G}{\rightarrow} S2$.

S_1: *Mexico City has a very bad pollution problem because the mountains around the city act as walls and block in dust and smog.*

S_2: *Poor air circulation out of the mountain-walled Mexico City aggravates pollution.*

To understand how TE by Generality can be modeled, we propose a new paradigm based on a new *Informative Asymmetric Measure* (IAM), called the *Asymmetric InfoSimba Similarity* (AIS) measure. Instead of relying on the exact matches of words between texts, we propose that one sentence entails the other by generality if two constraints hold: (a) if and only if both sentences share many related words and (b) if most of the words of a given sentence are more general than the words of the other one. As far as we know, we are the first to propose an unsupervised, language-independent, threshold-free methodology in the context of TEG.

In order to evaluate our methodology, it was necessary to create a corpus of pairs $T \rightarrow H$ and a set of TEG pairs ($T \overset{G}{\rightarrow} H$). This was achieved through the *CrowdFlower*[1] system, a convenient and fast way to collect annotations from a broad base of paid non-expert contributors over the Web. The corpus is composed of $T \rightarrow H$ pairs collected from the RTE challenge (RTE-1 through RTE-5). Only positive pairs of TE were submitted to CrowdFlower for annotation, together with a small set of carefully selected cases of known categorization that are used to train the participating annotators and to exercise quality control.

2. Corpus Construction

Large scale annotation projects such as TreeBank (Marcus et al., 1993), PropBank (Palmer et al., 2005), TimeBank (Pustejovsky et al., 2003), FrameNet (Baker et al., 1998), SemCor (Miller et al., 1993), and others play an important role in NLP research, encouraging the development of new ideas, tasks, and algorithms. The construction of these datasets, however, is extremely expensive in both annotator-hours and financial cost. Since the performance of many NLP tasks is limited by the amount and quality of data available to them (Banko and Brill, 2001), one promising alternative for some tasks is the collection of non-expert annotations. The availability and the increasing popularity of crowdsourcing services have been considered as an interesting opportunity to meet the aforementioned needs and design criteria.

Crowdsourcing services have been recently used with success for a variety of NLP applications (Callison-Burch and Dredze, 2010). Although MTurk is directly accessible only to US citizens, the CrowdFlower service provides a crowdsourcing interface to MTurk for non-US citizens.

The main idea in using crowdsourcing to create NLP resources is that the acquisition and annotation of large datasets needed to train and evaluate NLP tools and applications can be carried out in a cost-effective manner by defining simple Human Intelligence Tasks (HITs) routed to a crowd of non-expert workers, called *Turkers*, who are hired through online marketplaces.

2.1. Building Methodology - Quantitative Analysis

Our approach builds on a pipeline of HITs routed to MTurk workforce through the CrowdFlower interface. The objective is to collect $\langle T, H\rangle$ pairs where entailment by generality holds.

Our building methodology has several stages. First we select the positive pairs of TE from the first five RTE challenges. These pairs are then submitted to CrowdFlower through a job that we have built online, to be evaluated by Turkers. In CrowdFlower each $\langle T, H\rangle$ pair is a unit. The Turkers are asked to choose one of the following *Entailment by Generality* (TEG), *Entailment, but not Generality* (TEnG) or *Other*, whichever is most appropriate for the $\langle T, H\rangle$ pair under consideration.

Table 1 summarizes the work involved in the annotation of the entailment cases of the RTE-1 through RTE-5 datasets with the TEG, TEnG and *Other* labels. A total of 2,000 $\langle T, H\rangle$ pairs known to be in an entailment relation were uploaded, from which 1,740 were submitted for evaluation, and the remaining 260 constitute our *Gold* units.

[1] `http://crowdflower.com/` [Last access: 14th December, 2013]

	RTE-1	RTE-2	RTE-3	RTE-4	RTE-5
# Input Pairs[2]	400	400	400	500	300
# Pairs to Launch[3]	1,740				
# Gold Pairs[4]	260				
# Output Pairs[5]	1,203				
# Discarded Pairs[6]	797				
# *Trusted Turkers*	2,308				
# Trusted Judgments	5,220 (1,740*3)				
# Untrusted Judgments	60,482				
Evaluation Time	≈43 days				
Cost ($)	108.08				

Table 1: Summary of RTE by Generality corpus annotation task

In Table 1 we can see that 1,203 $\langle T, H \rangle$ pairs were annotated as TEG. Each pair was evaluated by three Turkers, and the final average inter-annotator agreement of 0.8 was verified.

This task proved to be hard for the Turkers, as it is difficult for human annotators to identify the entailment relation and entailment by generality in particular. This is proved by the time spent to complete the task (*Evaluation Time*) and the total number of *Judgments* (*Trusted* + *Untrusted*) needed to achieve the final objective.

The resulting manually annotated corpus is the first large-scale dataset containing a reasonable number of TEG pairs and constitutes one of the contributions of our work. It is an important resource available to the research community.

3. Asymmetric Association Measures

Most of the existing measures that evaluate the degree of similarity between words are symmetric (Pecina and Schlesinger, 2006; Tan et al., 2004). In order to avoid as much as possible the necessity of training data, different works propose the use of asymmetric association measures. Some have been introduced in the domain of taxonomy construction (Sanderson and Croft, 1999), others in cognitive psycholinguistics (Michelbacher et al., 2007) and in word order discovery (Dias et al., 2008).

Sanderson and Croft (1999) is one of the first studies to propose the use of *conditional probability* for taxonomy construction. They assume that a term t_2 subsumes a term t_1 if the documents in which t_1 occurs are a subset of the documents in which t_2 occurs constrained by $P(t_2|t_1) \geq 0.8$ and $P(t_1|t_2) < 1$. By gathering all subsumption relations, they build the semantic structure of any domain, which corresponds to a directed acyclic graph. In Sanderson and Lawrie (2000), the subsumption relation is relieved to the following expression $P(t_2|t_1) \geq P(t_1|t_2)$ and $P(t_2|t_1) > t$ where t is a given threshold and all term pairs found to have a subsumption relationship are passed through a transitivity module which removes extraneous subsumption relationships in such a way that transitivity is preferred over direct pathways, thus leading to a non-triangular directed acyclic graph.

In Michelbacher et al. (2007) the plain *conditional probability* and the *ranking measure* based on the Pearson's χ^2 test were used as a model for directed psychological association in the human mind. In particular, $R(t_2||t_1)$ returns the rank of t_2 in the association list of t_1 given by the order obtained with the Pearson's χ^2 test for all the words co-occurring with t_1. So, when comparing $R(t_2||t_1)$ and $R(t_1||t_2)$, the smaller rank indicates the strongest association.

In the specific domain of word order discovery, Dias et al. (2008) proposed a methodology combining directed graphs with the TextRank algorithm (Mihalcea and Tarau, 2004) to automatically induce a general-specific word order for a given vocabulary based on Web corpora frequency counts.

[2] Number of pairs $T \to H$ uploaded

[3] Number of pairs $T \to H$ submitted for evaluation

[4] Number of *Gold* pairs $T \to H$

[5] Number of pairs $T \to H$ classified as *Entailment by Generality*

[6] Number of pairs classified as *Entailment, but not Generality* or *Other*

In order to compute the general-specific relations between sentence pairs we have employed eight Asymmetric Association Measures (AAM) defined in the following equations: *Added Value* (Equation 1), *Braun-Blanket* (Equation 2), *Certainty Factor* (Equation 3), *Conviction* (Equation 4), *Gini Index* (Equation 5), *J-measure* (Equation 6), *Laplace* (Equation 7), and *Conditional Probability* (Equation 8).

$$AV(x\|y) = P(x|y) - P(x) \tag{1}$$

$$BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\overline{x},y)} \tag{2}$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)} \tag{3}$$

$$CO(x\|y) = \frac{P(x) \times P(\overline{y})}{P(x,\overline{y})} \tag{4}$$

$$GI(x\|y) = P(y) \times (P(x|y)^2 + P(\overline{x}|y)^2) - P(x)^2 \times P(\overline{y}) \times (P(x|\overline{y})^2 + P(\overline{x}|\overline{y})^2) - P(\overline{x})^2. \tag{5}$$

$$JM(x\|y) = P(x,y) \times \log\frac{P(x|y)}{P(x)} + P(\overline{x},y) \times \log\frac{P(\overline{x}|y)}{P(\overline{x})} \tag{6}$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2} \tag{7}$$

$$P(x|y) = \frac{P(x,y)}{P(y)} \tag{8}$$

3.1. Asymmetry between Sentences

There are a number of ways to compute the similarity between two sentences. Most similarity measures determine the distance between two vectors associated with two sentences (i.e. the vector space model). However, when applying the classical similarity measures between two sentences, only the identical indexes of the row vector X_i and X_j are taken into account, which may lead to miscalculated similarities. To deal with this problem, different methodologies have been proposed. A promising one is the InfoSimba informative similarity measure (Dias et al., 2007), expressed in Equation 9.

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl})}{\left(\begin{array}{l}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times S(W_{ik}, W_{il}) + \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times S(W_{jk}, W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl})\end{array}\right)}. \tag{9}$$

Here $S(.,.)$ is any symmetric similarity measure and each W_{ik} corresponds to the attribute word at the k^{th} position in the vector X_i, and p and q are the lengths of the vectors X_i and X_j respectively. This measure aims to compute the correlations between all pairs of words in two word context vectors instead of just relying on their exact match as with the cosine similarity measure. Furthermore, InfoSimba guarantees to capture similarity between pairs of sentences even when they do not share words. For example, this can happen when one sentence is a paraphrased version of the other and all the content words are substituted for similar words.

3.2. Asymmetric Similarities

Although there are many asymmetric similarity measures, they pose problems that may reduce their utility. On the one hand, asymmetric association measures can only evaluate the generality/specificity relation between words that are known to be in a semantic relation (Sanderson and Croft, 1999; Dias et al., 2008). Indeed, they generally capture the direction of association between two words based on document contexts and only take into account a loose semantic proximity between words. For example, it is highly probable to find that *Apple* is more general than *iPad*, which cannot be considered as a hypernymy/hyponymy or a meronymy/holonymy relation. On the other hand, asymmetric attributional word similarities only take into account common contexts to assess the degree of asymmetric relatedness between two words. To overcome this limitation, we introduce the *Asymmetric InfoSimba Similarity* measure (AIS), whose underlying idea is to say that one word x is semantically related to a word y and x is more general than y, if x and y share as many contexts as possible and each context word of x is likely to be more general than most of the context words of y. The AIS is defined in Equation 10, where $AS(.\|.)$ is any asymmetric word similarity measure, likewise for IS in Equation 9 where $S(.,.)$ stands for any symmetric similarity measure.

$$AIS(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl})}{\begin{pmatrix} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times AS(W_{ik}\|W_{il})+ \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times AS(W_{jk}\|W_{jl})- \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}) \end{pmatrix}}. \tag{10}$$

We now apply this idea to the RTEG problem, where each sentence is characterized by its content words and a sentence T is semantically related to sentence H and H is more general than T (i.e. $T \overset{G}{\rightarrow} H$), if H and T share as many related words as possible and each context word of H is likely to be more general than most of the words of T.

As a result, we propose a new simple and effective method for entailment identification through the AIS measure. We state that an entailment ($T \overset{G}{\rightarrow} H$) will hold if and only if $AIS(T\|H) < AIS(H\|T)$. Note that, contrarily to the existing methodologies, we do not need to define or tune any threshold at all. Indeed, due to its asymmetric definition, the *Asymmetric InfoSimba* similarity measure allows us to compare both sides of a candidate entailment.

Since we only want to compare $AIS(T\|H)$ and $AIS(H\|T)$, the denominator of AIS in both cases does not change. Thus we have defined an equivalent (with respect to the task) but simplified version of AIS – the $AISs(.\|.)$ in Equation 11, which ended up to be the one used in our experimentation.

$$AISs(X_i\|X_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}). \tag{11}$$

3.3. Three Levels of Word Granularity

It is evident that even through the simplified version of our proposed measure ($AISs$) we end up with a considerable amount of computation complexity – $O(n^2)$ – for comparing two sentences. Therefore, we have also considered two additional possibilities to reduce the number of words in each sentence without losing effectiveness. These are: (1) stop-word[7] removal and (2) multiword units (MWU) replacement, by identifying MWUs in the sentences. The MWUs were automatically computed using SENTA[8] (Dias et al., 1999) from the first five RTE datasets.

In summary, our experiments are based on three approaches to the calculations – using all words, using a list of stop words and finally using MWUs.

4. Experimentation and Results

In order to assess the effectiveness and general quality of our proposed measures for TEG identification, we have performed a comparative test on the corpus described in Section 2. We have tested our proposed $AISs$ measure with all word-similarity functions, mentioned in Section 3. Sentence similarity is computed in three different manners, as described previously in Section 3.3.

The evaluation functions used are based on the confusion matrix, in particular the accuracy and the precision. More specifically, we dealt with *Average Accuracy*, *Average Precision*, *Weighted Average Accuracy*, and *Weighted Average Precision*.

[7] A list of English stop-words, obtained using http://www.microsoft.com/en-us/download/confirmation.aspx?id=10024 [Last access: 14th December, 2013]

[8] The *Software for the Extraction of N-ary Textual Associations*.

4.1. The TEG Corpus

Here we report the obtained results of our methodology on the TEG corpus. These are the results we are most interested in as they concern the problem on which we are focusing our attention, namely identification of entailment by generality.

With respect to **Accuracy**, as seen in Table 2, the best performance, 0.85, is achieved by the measure *Braun-Blanket* in conjunction with the *MWU* method. The second best measure was *Added Value* with an accuracy of 0.69. It is important to highlight the significant difference between these two AAMs.

The measure *Braun-Blanket* remains the best one in the stop-word removal approach with an accuracy of 0.73, and *Gini Index* and *J-measure* achieved the second best results with an accuracy of 0.64. In *All Words* we have two measures with the best performance – *Conviction* and *J-measure* achieving respectively 0.70 and 0.69 of accuracy.

From Table 2 we may conclude that although *Conviction* is the best measure with *All Words* with respect to **Accuracy**, its performance is virtually equivalent to that of a random guesser for the *Without Stop Words* and *With MWU* approaches.

AAM	**Accuracy**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.67	0.63	0.69
BB	0.62	0.73	0.85
CF	0.65	0.63	0.64
P	0.61	0.60	0.64
CO	0.7	0.59	0.54
GI	0.65	0.64	0.68
JM	0.69	0.64	0.6
LP	0.64	0.62	0.6

Table 2: Accuracy by AAM

AAM	**Precision for A**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.81	0.74	0.82
BB	0.69	0.80	0.93
CF	0.74	0.67	0.63
P	0.72	0.70	0.64
CO	0.74	0.63	0.56
GI	0.74	0.72	0.65
JM	0.83	0.78	0.64
LP	0.71	0.69	0.58
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.45	0.47	0.49
BB	0.51	0.62	0.73
CF	0.51	0.56	0.68
P	0.45	0.44	0.63
CO	0.64	0.52	0.5
GI	0.51	0.51	0.71
JM	0.5	0.43	0.55
LP	0.52	0.51	0.63

Table 3: Precision by AAM

In terms of **Precision**, the *Braun-Blanket* measure, in conjunction with the *MWU* approach, achieved the best results for both entailment types: *Entailment by Generality* (**A**) and *Entailment, but no Generality* (**B**), with 0.93 and 0.73 points respectively. On **Precision A** the worst result is achieved by *Conviction* – 0.56 with *MWU*, and for **Precision A**, the worst result is achieved by *J-measure* with stop words removed: 0.43.

4.2. A Portuguese TEG Corpus

In this section we present the results of an experiment parallel to the one discussed in Section 4.1. The main idea was to measure the degree to which our methodology was capable to recognize TEGs in different languages. To this end, we have randomly selected a subset of 100 $\langle T, H \rangle$ pairs from the TEG Corpus, preserving the proportion of 60 $\langle T, H \rangle$ TEG pairs (Entailment by Generality) and 40 TEnG $\langle T, H \rangle$ pairs (Entailment, but no Generality). This subset of 100 TE pairs was translated into Portuguese using the *Google Translate* service.

Machine translation is a viable alternative to manual translation due to a combination of two factors. First, since our intention was to be as much language independent as possible, our methodology does not use morpho-syntactic analysis and language specific word order knowledge. On the other hand, *Google Translate* is reasonably successful in correct content word substitution. Thus, from the perspective of our bag-of-words approach *Google Translate* preserves well the important information. This supposition is in line with the fact that our results in Portuguese are comparable to the corresponding results in English.

With respect to **Accuracy** the best performance is achieved with the *Braun-Blanket* measure in conjunction with the *With MWU* approach, with a result of 0.76, as shown in Table 4. In this approach the second best measure is *Added Value*, with a result of 0.69. Similarly, *Braun-Blanket* achieves the best performance in the *Without Stop Words* approach, with a result of 0.71, followed by *Gini Index*, with 0.66. In *All Words*, the measure with the best **Accuracy** is *J-measure* (0.72).

In Table 4, the three measures with the lowest **Accuracy** are *Conditional Probability* in the approaches *All Words* and *Without Stop Words*, and *Conviction* – in *With MWU*.

AAM	**Accuracy**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.63	0.62	0.69
BB	0.62	0.71	0.76
CF	0.64	0.62	0.63
P	0.59	0.57	0.6
CO	0.68	0.6	0.5
GI	0.66	0.66	0.68
JM	0.72	0.58	0.6
LP	0.61	0.62	0.63

Table 4: Accuracy by AAM

Considering the **Accuracy** figures for English and for Portuguese, presented in Table 2 and Table 4, which show similar scale and variations, we conclude that the performance of our methodology is not significantly influenced by the language.

With respect to **Precision – Entailment by Generality** the measure *Braun-Blanket* in conjunction with the approach *With MWU* presents the best results (0.88), followed by *J-measure* in conjunction with the approach *All Words* (0.85). The worst results are achieved by *Certainty Factor* and *Laplace* in *With MWU* (0.6).

With respect to **Precision – Entailment, but no Generality** the results are markedly lower. The best results are achieved in *With MWU* by *Certainty Factor*, *Gini Index* and *Laplace* (0.68). The worst results are achieved by *Added Value* in *All Words* (0.38).

Both the **Accuracy** and the **Precision** figures show that whether applied to a corpus in English or in Portuguese, our methodology provides a classification capability that is significantly better than a random guessing baseline and virtually indistinguishable with respect to the language.

AAM	Precision for A		
	All Words	*Without Stop Words*	*With MWU*
AV	0.78	0.78	0.85
BB	0.65	0.78	0.88
CF	0.68	0.65	0.6
P	0.68	0.65	0.62
CO	0.73	0.65	0.55
GI	0.72	0.75	0.68
JM	0.85	0.72	0.62
LP	0.7	0.72	0.6
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.40	0.38	0.45
BB	0.58	0.6	0.58
CF	0.58	0.58	0.68
P	0.45	0.45	0.58
CO	0.60	0.52	0.43
GI	0.58	0.53	0.68
JM	0.53	0.38	0.58
LP	0.48	0.48	0.68

Table 5: Precision by AAM

5. Conclusion

This work presents a new methodology for recognizing TEG and studies its behavior in a detailed experimental configuration, achieving significant results. As seen in Table 2 and Table 3, there is always a measure and an approach that stand out, namely the *Braun-Blanket* measure in *With MWU*. However, *J-measure* and *Conviction* also have good results – (a) in **Precision – Entailment by Generality** *J-measure* with *All Words* has the second best performance (0.83) – in other words, *J-measure* with *All Words* has a good performance to identify entailment by generality between sentences; (b) *Conviction* ranks second for **Accuracy** (0.7) and achieves good results in **Precison – Entailment, but no generality or Other**, both in the *All Words* approach.

We may conclude that our methodology is language independent since results for Portuguese are comparable to those for English although with less significant discrimination between the first and the second measure. However, in terms of **Accuracy** (Table 4) and **Precision – Entailment by Generality** (Table 5) *Braun-Blanket* achieves the best performance in the approach *With MWU*.

With this paper we also contribute to the consideration of a new kind of textual entailment, providing also new experimental resources (TEG Corpus). Our methodology is unsupervised and language independent, and accounts for the asymmetry of the studied phenomena by means of asymmetric similarity measures. Using our methodology we have demonstrated excellent results in identifying textual entailment by generality.

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL 1998)*, volume 1, pages 86–90.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, pages 26–33.

Callison-Burch, C. and Dredze, M. (2010). Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 1–12.

Dagan, I., Glickman, O., and Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005*, pages 177–190.

Dias, G., Guilloré, S., and Lopes, J. (1999). Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. In *Proceedings of the 6ème Confrence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, pages 333–339.

Dias, G., Alves, E., and Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, pages 1334–1340.

Dias, G., Mukelov, R., and Cleuziou, G. (2008). Unsupervised Graph-Based Discovery of General-Specific Noun Relationships from Web Corpora Frequency Counts. In *Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008)*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Michelbacher, L., Evert, S., and Schtze, H. (2007). Asymmetric Association Measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, pages 1–6.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT 1993, pages 303–308.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Pecina, P. and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006)*, pages 651–658.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

Sanderson, M. and Croft, B. (1999). Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 206–213.

Sanderson, M. and Lawrie, D. (2000). Building, Testing, and Applying Concept Hierarchies. *Advances in Information Retrieval*, 7:235–266.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the Right Objective Measure for Association Analysis. *Information Systems*, 29(4):293–313.