Textual Entailment by Generality

Gaël Dias and Sebastião Pais
University of Beira Interior, Portugal
{ddg,sebastiao}@hultig.di.ubi.pt

Katarzyna Wegrzyn-Wolska
ESIGETEL, France
katarzyna.wegrzyn@esigetel.fr

Robert Mahl
Ecole Nationale Supérieure des Mines de Paris, France
mahl@ensmp.fr

Abstract

Textual Entailment consists in determining if an entailment relation exists between two texts. In this paper, we present an Informative Asymmetric Measure called the Simplified Asymmetric InfoSimba ($AISs$), which we combine with different asymmetric association measures to recognize the specific case of Textual Entailment by Generality. In particular, the $AISs$ proposes an unsupervised, language-independent, threshold free solution. This new measure is tested against the first Recognizing Textual Entailment data set for an exhaustive number of asymmetric association measures and shows that the combination of the $AISs$ with the Braun-Blanket steadily improves results.

1 Introduction

Recognizing Textual Entailment is a key task for many natural language processing (NLP) problems. It consists in determining if an entailment relation exists between two texts: the text *T* and the hypothesis *H*. The notation $T \rightarrow H$ says that the meaning of *H* can be inferred from *T*. More formally, a textual entailment is defined as a directional relation between the entailing text *T* and the entailed text *H*. It is then said that *T* entails *H* if, typically, a human reading *T* would infer that *H* is most likely true based on the truth of *T*. In this paper, we introduce the paradigm of Textual Entailment by Generality, which can be defined as the entailment from a specific sentence towards a more general sentence. For example, from sentences (1) and (2) extracted from RTE-1[1], we would easily state that $(1) \rightarrow (2)$ as their meaning is roughly the same although sentence (2) is more general than sentence (1).

To understand how Textual Entailment by Generality can be modeled for two sentences, we propose a new paradigm based a new informative asymmetric measure, called the Simplified Asymmetric InfoSimba similarity measure ($AISs$). Instead of relying on the exact matches of words between texts, we propose that one sentence infers the other one in terms of generality if two constraints are respected: (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 sentence. As far as we know, we are the first to propose an unsupervised, language-independent, threshold free methodology in the context of Textual Entailment by Generality, although the approach from [1] is based on similar assumptions.

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

(2) *Poor air circulation out of the mountain-walled Mexico City aggravates pollution.*

This new proposal is exhaustively evaluated against the RTE-1 data set by testing different asymmetric association measures in combination with the $AISs$. In particular, we chose the RTE-1[2] as it is the only data set for which there exist comparable results with linguistic free methodologies [1] [2] [3]. The evaluation shows promising results and evidences that the combination of the $AISs$ with the Braun-Blanket steadily improves results.

2 Related Works in RTE-1

Different approaches have been proposed to recognize Textual Entailment: from unsupervised language-independent methodologies [1] [2] [3] to deep linguistic analyses [4] [5] [6]. In this section, we will particularly mention the unsupervised language-independent approaches, which can be directly compared to our proposal, at least to a certain extent.

[1] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/

[2] This year will be RTE-7 Challenge.

One of the most simple proposal is the one proposed by [2] who explore the BLEU algorithm. First, for several values of n (typically from 1 to 4), they calculate the percentage of n-grams from the text T, which appear in the hypothesis H. The frequency of each n-gram is limited to the maximum frequency with which it appears in any text T. Then, they combine the marks obtained for each value of n, as a weighted linear average and finally apply a brevity factor to penalize short texts T. The output of BLEU is then taken as the confidence score. Finally, they perform an optimization procedure to choose the best threshold according to the percentage of success of correctly recognized entailments. The value obtained was 0.157. Thus, if the BLEU output is higher than 0.157, the entailment is marked as true, otherwise as false.

A second more interesting work is proposed by [3], where the entailment data is treated as an aligned translation corpus. In particular, they use the GIZA++ toolkit to induce alignment models. However, the alignment scores alone were next to useless for the RTE-1 development data, predicting entailment correctly only slightly above chance. As a consequence, they introduced a combination of metrics intended to measure translation quality. All but one of these metrics come from libparis, a library of string similarity metrics assembled by MITRE. Finally, they combined all the alignment information and string metrics with the classical K-NN classifier to choose, for each test pair, the dominant truth value among the five nearest neighbors in the development set.

The most interesting work is certainly the one described in [1] who propose a general probabilistic setting that formalizes the notion of Textual Entailment. Here, they focus on identifying when the lexical elements of a textual hypothesis H are inferred from a given text T. The lexical entailment probability is derived from Equation 1 where $hits(.)$ is a function that returns the number of documents, which contain its arguments.

$$P(H|T) = \prod_{u \in H} max_{v \in T} \frac{hits(u,v)}{hits(v)} \tag{1}$$

The text and hypothesis of all pairs in the development and test sets were tokenized and stop words were removed to empirically tune a decision threshold, λ. So, for a pair $T - H$, they tagged an example as true (i.e. entailment holds) if $P(H|T) > \lambda$, and as false otherwise. The threshold was set to 0.005 for best performance.

The best results from these three approaches are obtained by [1], who introduce the notion of asymmetry within their model without clearly mentioning it. The underlying idea is based on the fact that for each word in H, the best asymmetrically co-occurring word in T is chosen to evaluate $P(H|T)$. Although all three approaches show interesting properties, they all depend on tuned thresholds, which can not reliably be reproduced and need to be changed for each new applications. Moreover, they need training data, which may not be available. Our idea aims at generalizing the hypothesis made by [1]. Indeed, their methodology is only based on one pair $(u, v), \forall u$ and does not take into account the fact that many pairs i.e. $(u, v), \exists v \forall u$ may help the decision process. Moreover, they do not propose a solution for the case where the ratio $\frac{hits(u,v)}{hits(v)}$ is null. Finally, we propose to avoid the definition of a "hard" threshold and study exhaustively asymmetry in language i.e. not just by the conditional probability as done in [1]. For that purpose, we propose a new Informative Asymmetric Measure called the Simplified Asymmetric InfoSimba ($AISs$) combined with different Association Measures.

3 Asymmetry between Words

New trends have recently emerged with the study of asymmetric measures [9]. The idea of an asymmetric measure is inspired by the fact that within the human mind, the association between two words or concepts is not always symmetric. Within this scope, seldom new researches have been emerging, which we believe can lead to great improvements in the field of NLP. In order to keep language-independency and to some extent propose unsupervised methodologies, different works have been proposing the use of asymmetric association measures [7] [8]. Here, we present eight asymmetric association measures that will be tested: Conditional Probability (Equation 2), Added Value (Equation 3), Braun-Blanket (Equation 4), Certainty Factor (Equation 5), Conviction (Equation 6), Gini Index (Equation 7), J-measure (Equation 8) and Laplace (Equation 9).

$$P(x|y) = \frac{P(x,y)}{P(y)}. \tag{2}$$

$$AV(x\|y) = P(x|y) - P(x). \tag{3}$$

$$BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\bar{x},y)}. \tag{4}$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)}. \tag{5}$$

$$CO(x\|y) = \frac{P(x) \times P(\bar{y})}{P(x,\bar{y})}. \tag{6}$$

$$\begin{aligned} GI(x\|y) = P(y)(P(x|y)^2 + P(\bar{x}|y)^2) - P(x)^2 + \\ P(\bar{y})(P(x|\bar{y})^2 + P(\bar{x}|\bar{y})^2) - P(\bar{x})^2. \end{aligned} \tag{7}$$

$$JM(x\|y) = P(x,y) \times \log \frac{P(x|y)}{P(x)} + P(\bar{x},y) \times \log \frac{P(\bar{x}|y)}{P(\bar{x})}. \tag{8}$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2}. \tag{9}$$

4 Asymmetry between Sentences

Although there are many asymmetric similarity measures between words, there does not exist any attributional similarity measure capable to assess whether a sentence is more specific/general than another one. To overcome this issue, we introduce the Simplified Asymmetric InfoSimba similarity measure ($AISs$), which underlying idea is to say that a sentence T is semantically related to sentence H and H is more general than T (i.e. $T \rightarrow H$), if H and T share as many relevant related words as possible between contexts and each context word of H is likely to be more general than most of the context words of T. The $AISs$ is defined in Equation 10, where $AS(.\|.)$ is any asymmetric similarity measure between two words introduced in section 3.

$$AISs(X_i\|X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}\|W_{jl}). \quad (10)$$

As a consequence, an entailment ($T \rightarrow H$) will hold if and only if $AISs(T\|H) < AISs(H\|T)$. Otherwise, the entailment will not hold. This way, unlike existing methodologies, we do not need to define or tune any threshold. Indeed, due to its asymmetric definition, the $AISs$ allows to compare both sides of entailments.

5 Results and Discussion

In order to perform our evaluation, we ran our methodology over the RTE-1 data set, which contains seven tasks: CD (Comparable Documents), IE (Information Extraction), MT (Machine Translation), PP (Paraphrases), QA (Question Answering), RC (Reading Compression) and IR (Information Retrieval). We show the accuracy results for all data sets individually and the global accuracy for all the unsupervised language-independent methodologies in Tables 1 and 2. In particular, we used the Google API to calculate all joint and marginal frequencies. So, instead of relying on a closed corpus and exact frequencies, we based our analysis on the Web and Web hits i.e. estimated number of documents where words appear. We also compare our methodology with both [1] and [2]. Unfortunately, the data for [3] were not available but from the RTE-1 challenge we know that it performed worst than [1].

On average, [1] shows the best results with 57% accuracy compared to the combination of the $AISs$ with the Braun-Blanket, which reaches 54%. In terms of the overall RTE-1 challenge, we would take the sixth place just after [1]. However, when analyzing the results of [1] in more detail, we clearly see that the good figures are mainly obtained due to very high accuracy for the CD data set compared to the other ones. Indeed, we show that we overtake [1] for the IE, PP, QA and RC collections with the Braun-Blanket as well as for the IR collection with the Added Value. These results are particularly interesting as they show that the conditional probability alone may not be a good indicator to tackle specific entailments. Indeed, it shows low levels of accuracy for many tasks. Another important result is the fact that we would achieve 55% accuracy by taking the best measures for each one of the collections. At this point of our evaluation, it is important to point at that the fact that we do not remove words from stop-lists unlike in [1]. This is a major difference because if [1] used plain raw texts, results may be lower, especially based on the fact that they use a product of conditional probabilities. The comparative results of [1] with and without stop words are presented in Table 3 and show that in this case, the maximum obtained accuracy

Table 1. Accuracy by Data Set (1).

	CD	IE	MT	PP
Glickman et al.[1]	**0.83**	0.56	**0.57**	0.52
Added Value	0.49	0.53	0.53	0.60
J-measure	0.46	0.52	0.46	**0.62**
Braun-Blanket	0.47	**0.57**	0.51	**0.62**
Laplace	0.49	0.52	0.53	0.54
Perez et al.[2]	0.7	0.50	0.38	0.46
Certainty Factor	0.46	0.56	0.53	0.52
Conditional Probability	0.49	0.52	0.53	0.54
Gini Index	0.47	0.48	0.48	0.40
Conviction	0.47	0.46	0.55	0.48

Table 2. Accuracy by Data Set (2).

	QA	RC	IR	ALL
Glickman et al.[1]	0.49	0.53	0.50	**0.57**
Added Value	0.49	0.51	**0.56**	0.53
J-measure	0.52	0.52	0.53	0.52
Braun-Blanket	**0.54**	**0.54**	0.54	0.54
Laplace	0.50	0.50	0.48	0.50
Perez et al.[2]	0.42	0.46	0.49	0.49
Certainty Factor	0.50	0.51	0.48	0.51
Conditional Probability	0.50	0.50	0.49	0.51
Gini Index	0.53	0.46	0.50	0.47
Conviction	0.49	0.50	0.38	0.48

Table 3. [1] with and without stop words.

	CD	IE	MT	PP
[1] with stop words	0.53	0.51	0.63	0.44
[1] without stop words	0.83	0.56	0.57	0.52
	QA	**RC**	**IR**	**ALL**
[1] with stop words	0.54	0.48	0.54	0.52
[1] without stop words	0.49	0.53	0.50	0.57

would be 52%, i.e. above our 54% accuracy result. In this aspect, we provide a more universal solution (and as such really language-independent) capable of handling raw texts comparatively to most other methodologies.

To understand better the results, we decided to look at the precisions for entailment and no entailment individually for the best two measures i.e. Braun-Blanket and Added Value. These results are available in Tables 4 and 5. They clearly show that recognizing Textual Entailment is a difficult task as levels just over chance are obtained. However, there are some tasks such as PP and IR for which promising results are obtained. But there are also some tasks, which revealed difficult to solve such as CD and QA. Maybe the most interesting results from Tables 4 and 5 is the fact that they show comparable precisions between entailment and no entailment, which shows that the proposed methodology is robust and that the confusion matrix is balanced.

Table 4. Precision for Entailment.

	CD	IE	MT	PP
Added Value	0.49	0.53	0.53	0.65
Braun-Blanket	0.46	0.58	0.51	0.71
	QA	**RC**	**IR**	**ALL**
Added Value	0.48	0.51	0.56	0.54
Braun-Blanket	0.54	0.54	0.58	0.56

Table 5. Precision for No Entailment.

	CD	IE	MT	PP
Added Value	0.49	0.52	0.53	0.58
Braun-Blanket	0.47	0.56	0.51	0.58
	QA	**RC**	**IR**	**ALL**
Added Value	0.48	0.51	0.55	0.52
Braun-Blanket	0.54	0.53	0.53	0.53

6 Conclusion

In this paper, we proposed a new unsupervised, language-independent, threshold free methodology to recognize Textual Entailment by Generality. For that purpose, we proposed a new attributional similarity measure, called the Simplified Asymmetric InfoSimba similarity measure ($AISs$), capable of assessing whether a sentence is more specific/general than another one. To test our hypothesis, we evaluated our model based on eight asymmetric association measures over the RTE-1 data collection. The results show promising results as we obtain better results than [1] when keeping stop-words. Moreover, the results of [1] are highly over-evaluated based on their results for the CD collection. However, this first exploratory study opens a great deal of new research directions. In particular, we need (1) to assess why the obtained results are so low for the CD collection comparatively to other methodologies and (2) why the Braun-Blanket measure seems to better adapt to the RTE-1 tasks, as this situation can easily be explained for the Added Value.

References

[1] Glickman, O. and Dagan, I. and Koppel, M. 2005. *Web Based Probabilistic Textual Entailment*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 33-36, 11-13 April, Southampton, U.K.

[2] Pérez, D. and Alfonseca, E. 2005. *Application of the Bleu Algoritm for Recognizing Textual Entailments*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 9-12, 11-13 April, Southampton, U.K.

[3] Bayer, S. and Burger, J. and Ferro, L. and Henderson, J. and Yeh, A. 2005. *MITRE's submissions to the EU Pascal RTE Challenge*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 41-44, 11-13 April, Southampton, U.K.

[4] Newman, E. and Stokes, N. and Dunnion, J. and Carthy, J. 2005. *UCD IIRG Approach to the Textual Entailment Challenge*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 53-56, 11-13 April, Southampton.

[5] Delmonte, R. and Tonelli, S. and Boniforti, M. and Bristot, A. and Pianta, E. 2005. *VENSES - a Linguistically-Based System for Semantic Evaluation*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 49-52, 11-13 April, Southampton, U.K.

[6] Herrera, J. and Peñas, A. and Verdejo, F. 2005. *Textual Entailemnt Recognition Based on Dependency Analysis and WordNet*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 21-24, 11-13 April, Southampton.

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

[8] Tan, P-N. and Kumar, V. and Srivastava, J. 2005. *Selecting the Right Objective Measure for Association Analysis*. Information Systems, 29, 4, 293-313.

[9] Michelbacher, L. and Evert, S. and Schütze, H. 2007. *Asymmetric Association Measures*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007). 1-6.