Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 00 (2011) 1–9

Procedia Social and Behavioral Sciences

www.elsevier.com/locate/procedia

12th International Conference of the Pacific Association For Computational Linguistics

Textual Entailment by Generality

Gaël Dias[a,b], Sebastião Pais[a,d], Katarzyna Wegrzyn-Wolska[c], Robert Mahl[d], Pierre Jouvelot[d]

[a]*HULTIG - University of Beira Interior, Portugal*
[b]*DLU/GREYC - University of Caen Basse-Normandie, France*
[c]*SITR - École Supérieure d'Ingénieurs en Informatique et Génie des Télécommunications, France*
[d]*CRI - École Nationale Supérieure des Mines de Paris, France*

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 Asymmetric InfoSimba (*AIS*), which we combine with different asymmetric association measures to recognize the specific case of Textual Entailment by Generality. In particular, the *AIS* proposes an unsupervised, language-independent, threshold free solution. This new measure is tested against the first Recognizing Textual Entailment dataset for an exhaustive number of asymmetric association measures and shows that the combination of the *AIS* with the Braun-Blanket steadily improves results against competitive measures such as the one proposed by [1].

Keywords: Asymmetric Association Measures; Informative Asymmetric Measure; Textual Entailment

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).

Email addresses: ddg@hultig.di.ubi.pt (Gaël Dias), sebastiao@hultig.di.ubi.pt (Sebastião Pais), katarzyna.wegrzyn@esigetel.fr (Katarzyna Wegrzyn-Wolska), mahl@ensmp.fr (Robert Mahl), pierre.jouvelot@mines-paristech.fr (Pierre Jouvelot)

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

(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.*

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 Asymmetric InfoSimba similarity measure (AIS). 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. This new proposal is exhaustively evaluated against the RTE-1 dataset by testing different asymmetric association measures in combination with the AIS. In particular, we chose the RTE-1[2] as it is the only dataset 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 AIS with the Braun-Blanket steadily improves results against competitive measures such as the one proposed by [1].

2. Recognizing Textual Entailment

The detection of Textual Entailment is a recently recognized challenge in the NLP domain and one of the most demanding. Indeed, participating systems must prove their capabilities of understanding how language works. Indeed, as stated in [4], recognizing entailment bears similarities to Turing's famous test to assess whether machines can think, as access to different sources of knowledge and the ability to draw inferences seem to be among the primary ingredients for an intelligent system. Moreover, many NLP tasks have strong links to entailment. In particular, the RTE-1 challenge listed the following ones. In Summarization (SUM), a summary should be entailed by the text. Paraphrases (PP) can be seen as mutual entailment between a text T and a hypothesis H. In Information Extraction (IE), the extracted information should be entailed by the text content whereas in Machine Translation (MT), the automatic translation should at most be entailed by the golden standard human translation. In Comparable Document (CD) analysis, annotators should be able to identify $T \rightarrow H$ pairs, for example, by examining a cluster of comparable news articles that cover a common story. In Reading Comprehension (RC), a typical reading comprehension exercise in human language teaching, students are asked to judge whether a particular assertion can be inferred from a given text story and in Question Answering (QA), the answer obtained for one question after the Information Retrieval process must be entailed by the supporting text. The RTE-1 Challenge started in 2005 [5] as an attempt to promote an abstract generic task that captures major semantic inference needs across applications. Participants are provided with pairs of short texts, which are called Text-Hypothesis ($T \rightarrow H$) to learn or tune their models. Then, a test dataset is provided to run the systems and the performances are gathered. The collected examples represent a range of different levels of entailment reasoning, based on lexical, syntactic, logical and world knowledge at different levels of difficulty. To have an idea of the kind of entailments proposed in each dataset, we show positive and negative examples of textual entailments.

Comparable Document (CD):

<pair id="754" value="TRUE" task="CD">
<t>Mexico City has a very bad pollution problem because the mountains around the city act as walls and block in dust and smog.</t>
<h>Poor air circulation out of the mountain-walled Mexico City aggravates pollution.</h></pair>
<pair id="755" value="FALSE" task="CD">
<t>The Alameda Central (Central Mall) roughly marks the western edge of the old colonial city.</t>
<h>The Alameda Central is west of the Zocalo.</h></pair>

[2] This year will hold the RTE-7 Challenge.

Information Extraction (IE):

<pair id="1697" value="TRUE" task="IE">
<t>In 1999 Ford bought the apartment in Manhattan that he shares with his new girlfriend, Calista Flockhart.</t>
<h>Calista Flockhart lives in Manhattan.</h></pair>
<pair id="1650" value="FALSE" task="IE">
<t>Canadian authorities are investigating a report that Frank Costello, an American citizen, was offered the job of assassinating Prime Minister Brian Mulroney, a spokeswoman said.</t>
<h>Frank Costello killed Prime Minister Brian Mulroney.</h></pair>

Machine Translation (MT):

<pair id="1216" value="TRUE" task="MT">
<t>The airport will not be opened without the go-ahead from the Palestinian Authority in Gaza, at which point the Moroccan delegation will oversee the running of the airport.</t>
<h>A Moroccan delegation is waiting for the green light from the Palestinian Authority to supervise the airport's operation.</h></pair>
<pair id="1325" value="FALSE" task="MT">
<t>In turn, the Editor-in-Chief of Al Jumhoria Newspaper was appointed Ambassador of Iraq to India.</t>
<h>Al Jumhoria is the Iraqi Ambassador to India.</h></pair>

Paraphrase Acquisition (PP):

<pair id="2049" value="TRUE" task="PP">
<t>Five other soldiers have been ordered to face courts-martial.</t>
<h>Five other soldiers have been demanded to face courts-martial.</h></pair>
<pair id="1984" value="FALSE" task="PP">
<t>Those accounts were not officially confirmed by the Uzbek or American governments.</t>
<h>The Uzbek or American governments confirmed those accounts.</h></pair>

Question Answering (QA):

<pair id="1595" value="TRUE" task="QA">
<t>In 1884, with Ireland under the rule of the British Crown, a group of Irish nationalists met in County Galway to establish an organization for Irish athletes, the Gaelic Athletic Association (GAA).</t>
<h>The GAA was set up in Co. Galway.</h></pair>
<pair id="2029" value="FALSE" task="QA">
<t>Allen Overy has a leading practice in both France and Europe and the firm's partners devoted to capital markets work in Paris are principally Dan Lauder and Francois Poudelet.</t>
<h>Paris is the capital of France.</h></pair>

Reading Compression (RC):

<pair id="1082" value="TRUE" task="RC">
<t>The tests cover seven subject areas and are given annually.</t>
<h>The tests are given once a year.</h></pair>
<pair id="2126" value="FALSE" task="RC">
<t>One person was also killed and four wounded when a mortar round hit a hospital.</t>
<h>A hospital containing four wounded people killed one when a mortar hit.</h></pair>

Information Retrieval (IR):

<pair id="298" value="TRUE" task="IR"> <t>Italy and Germany have each played twice, and they haven't beaten anybody yet.</t>
<h>Neither Italy nor Germany have won yet.</h></pair>
<pair id="2025" value="FALSE" task="IR">
<t>There are a lot of farmers in Poland who worry about their future if Poland joins the European Union.</t>
<h>Poland joins the European Union.</h></pair>

3. 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 [6, 7, 8]. In this section, we will particularly detail the unsupervised language-independent approaches, which can be directly compared to our proposal, at least to a certain extent. One of the most simple proposal is the one proposed by [2] who explore the BLEU algorithm [9]. 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 [10] 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.

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

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 Asymmetric InfoSimba (AIS) combined with different Association Measures.

4. Asymmetry between Words

Most of the metrics, which evaluate the degree of similarity between words are symmetric [11, 12], except perhaps pattern-based similarities [13, 14]. Patterns can be helpful to learn knowledge from texts that can possibly be expressed by constructions known in advance and surely embody the easiest way to induce this knowledge. Most of the works in this area have been dealing with the identification of the hypernymy/hyponymy relation although some other word semantic relations such as synonymy and

meronymy/holonymy have been tackled. In order to extract hypernymy/hyponymy relations, [13] first identifies a set of lexical-syntactic patterns that are easily recognizable (i.e. occur frequently and across text genre boundaries). These can be called seed patterns. Based on these seeds, she proposes a bootstrapping algorithm to semi-automatically acquire new more specific patterns such as *such NP as (NP,)* {or | and} NP*. Similarly, [14] uses predefined patterns such as *X is a (kind of) Y* or *X, Y, and other Zs*, following the discussion in [15] that nouns in conjunctions or appositive relations tend to be semantically related. Despite the variety of approaches, two common characteristics are transversal to the methodology: (1) the necessity of manual effort as to compose the patterns and (2) the language-dependency of the method. Other drawbacks can be identified. In particular, lexical-syntactic patterns tend to be quite ambiguous as to which relations they indicate and this worsens when ambiguous words are involved. Also, mainly subsets of possible instances of semantic relations are likely to appear, thus imposing the existence of a great number of seed patterns. To overcome such drawbacks, new trends have recently emerged with the study of asymmetric measures [16]. 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. For example, as stated in [16], "*there is a tendency for a strong forward association from a specific term like adenocarcinoma to the more general term cancer, whereas the association from cancer to adenocarcinoma is weak*". For instance, *cancer* would be more central than *adenocarcinoma*. Within this scope, seldom new researches have been emerging over the past few years, which propose the use of asymmetric similarity measures, which we believe can lead to great improvements in the acquisition of word semantic relations as shown in [18]. In order to keep language-independency and to some extent propose unsupervised methodologies, different works propose to use asymmetric association measures. Some have been introduced in the domain of taxonomy construction [19], others in cognitive psycholinguistics [16] and in word order discovery [17]. In the domain of taxonomy construction, [19] is certainly one of the first studies to propose the use of the conditional probability as defined in Equation 2.

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

Later, [16] proposed two different measures to model the notion of asymmetric association. Their intent is to determine to what extent these two measures of directed association can be used as a model for directed psychological association in the human mind. These two measures are the plain conditional probability and the ranking measure $R(.\|.)$ based on the Pearson's χ^2 test. 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. Finally, in the specific domain of word order discovery, [17] proposed a methodology based on directed graphs and the TextRank algorithm [20] to automatically induce a general-specific word order for a given vocabulary based on Web corpora frequency counts. For that purpose, they used eight different asymmetric association measures to build an asymmetric word-word matrix for a given vocabulary. A directed graph is obtained by keeping the edges, which correspond to the maximum value of the asymmetric association measure between two words. We present the eight asymmetric association measures used in this work that will be evaluated in the context of asymmetry between sentences: the Added Value (Equation 3), the Braun-Blanket (Equation 4), the Certainty Factor (Equation 5), the Conviction (Equation 6), the Gini Index (Equation 7), the J-measure (Equation 8), the Laplace (Equation 9), and the Conditional Probability (Equation 2).

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

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

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

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

$$GI(x||y) = P(y)(P(x|y)^2 + P(\bar{x}|y)^2) - P(x)^2 + P(\bar{y})(P(x|\bar{y})^2 + P(\bar{x}|\bar{y})^2) - P(\bar{x})^2. \quad (7)$$

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

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

5. 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 to 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, but the most promising one is certainly the one proposed by [21], the InfoSimba informative similarity measure, expressed in Equation 10 where $S(.,.)$ is any symmetric similarity measure, each W_{ij} corresponds to the attribute word at the j^{th} position in the vector X_i and X_{ik} is the weight of word W_{ik}.

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.S(W_{ik}, W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.S(W_{ik}, W_{il}) + \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.S(W_{jk}, W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.S(W_{ik}, W_{jl}) \end{array}\right)}. \quad (10)$$

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 asymmetric InfoSimba similarity measure (AIS), 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 AIS is defined in Equation 11, where $AS(.||.)$ is any asymmetric similarity measure between two words introduced in section 4.

$$AIS(X_i||X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}||W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.AS(W_{ik}||W_{il}) + \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jk}.AS(W_{jk}||W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}||W_{jl}) \end{array}\right)}. \quad (11)$$

As computation of the AIS may be hard due to orders of complexity, we also define its simplified version $AIS\,s(.||.)$ in Equation 12, which we will specifically use in our experiments.

$$AIS\,s(X_i||X_j) = \frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.AS(W_{ik}||W_{jl}). \quad (12)$$

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

6. Results and Discussion

In order to perform our evaluation, we ran our methodology over the RTE-1 dataset, 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 Web hits i.e. estimated number of documents where words appear. We also compare our methodology with both [1] and [2]. Unfortunately, the results for [3] were not available but from the RTE-1 challenge we know that it performed worst than [1].

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]	0.70	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

On average, [1] shows the best results with 57% accuracy compared to the combination of the AIS s 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 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.

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

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

7. Conclusions

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 Asymmetric InfoSimba similarity measure (*AIS*), 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 showed promising results as we obtained 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 as it takes into account the probability to encounter a context word by chance.

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] Bos, J. and Markert, K. 2005. *Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 65-68, 11-13 April, Southampton, U.K.

[5] Dagan, I. and Glickman, O. and Magnini, B., 2005. *The PASCAL Recognising Textual Entailment Challenge*, Proceedings of the First Challenge Workshop Recognising Textual Entailment, 11-13 April, Southampton, U.K.

[6] 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, U.K.

[7] 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.

[8] 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, U.K.

[9] Papineni, K. and Roukos, S. and Ward, T. and Zhu, W. 2001. *BLEU: a Method for Automatic Evaluation of Machine Translation*, Research report, IBM.

[10] Och, F.J. and Ney, H. 2003. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, 29(1), 19-51.

[11] 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.

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

[13] Hearst, A. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. Proceedings of the 14th Conference on Computational linguistics (COLING 1992), 539-545.

[14] Caraballo, A. 1999. *Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999), 120-126.

[15] Riloff, E. and Shepherd, J. 1997. *A Corpus-Based Approach for Building Semantic Lexicons*. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), 117-124.

[16] 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.

[17] Dias, G. and Mukelov, R. 2008. *Unsupervised Graph-Based Discovery of General-Specific Noun Relationships from Web Corpora Frequency Counts*. Proceedings of the 12th International Conference on Natural Language Learning (CONLL 2008). 147-153.

[18] Cleuziou, G. and Dias, G. and Buscaldi, D. and Levorato, V. 2011. *A Pretopological Framework for Automatic Construction of Semantic-Lexical Structures from Texts*. Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011).

[19] Sanderson, M. and Croft, B. 1999. *Deriving Concept Hierarchies from Text*. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1999), 206-213.

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

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