# Recognizing Textual Entailment by Generality using Informative Asymmetric Measures and Multiword Unit Identification to Summarize Ephemeral Clusters

Gaël Dias*†, Sebastião Pais*, Katarzyna Wegrzyn-Wolska‡ and Robert Mahl§
**HULTIG*
*University of Beira Interior, Covilhã, Portugal*
*Email: {ddg, sebastiao}@hultig.di.ubi.pt*
†*DLU - GREYC*
*University of Caen Basse-Normandie, Caen, France*
‡*ESIGETEL*
*Fontainebleau, France*
*Email: katarzyna.wegrzyn@esigetel.fr*
§*ENSMP*
*Paris, France*
*Email: mahl@ensmp.fr*

***Abstract*—In the context of Ephemeral Clustering of web Pages, it can be interesting to label each cluster with a small summary instead of just a label. Within this scope, we introduce the paradigm of Textual Entailment by Generality, which can be defined as the entailment from a specific web snippet towards a more general web snippet. The subjacent idea is to find the best web snippet, which summarizes and subsumes all the other web snippets within an ephemeral cluster. To reach this objective, we first propose a new informative asymmetric similarity measure called the Simplified Asymmetric InfoSimba ($AISs$), which can be combined with different asymmetric association measures. In particular, the $AISs$ proposes an unsupervised language-independent solution to infer Textual Entailment by Generality and as such can help to encounter the web snippet with maximum semantic coverage. This new methodology is tested against the first Recognizing Textual Entailment data set (RTE-1)[1] for an exhaustive number of asymmetric association measures with and without the identification of Multiword Units. The comparative experiments with existing state-of-the-art methodologies show promising results.**

***Keywords*-Asymmetric Association Measures; Informative Asymmetric Measure; Multiword Units Identification; Textual Entailment by Generality**

## I. INTRODUCTION

Although, Ephemeral Clustering has been studied for more than a decade, it has received low user acceptance. According to us, there are two main reasons for this situation. First, state-of-the-art systems tend to generate an excessive number of clusters. As a consequence, browsing through a high number of clusters is mostly similar to searching through a high number of Web pages. Second, improved user interfaces can only be achieved through high quality cluster labeling. In the optimal case, the labels of the clusters should clearly evidence their overall contents. However, very little has been proposed in the community to overcome the latter situation. The only exception is certainly [1] who propose to increase the expressiveness of label clusters with a summary obtained by classical Multi-document Summarization techniques. However, their solution is full-text based and can not be applied in real-time real-world applications. As a consequence, we propose to increase cluster expressiveness based on finding the web snippet within the ephemeral cluster, which best summarizes and subsumes all the other web snippets present in the cluster. For that purpose, we propose a different methodology based on Textual Entailment.

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*. In this paper, we introduce the paradigm of Textual Entailment by Generality, which can be defined as the entailment from a specific web snippet towards a more general web snippet. The idea behind is to find the best web snippet, which summarizes and subsumes all the other web snippets within an ephemeral cluster and as such can be defined as a good cluster candidate summary.

However, before reaching this step, we need to understand how Textual Entailment by Generality can be modeled for two sentences. For that purpose, we propose a new paradigm based a new informative asymmetric measure, called the Asymmetric InfoSimba similarity measure ($AISs$). So, instead of relying on the exact matches of words between texts, we propose that one sentence/snippet infers the other one in terms of generality when (a) if and only if both sentences/snippets share a great content of related words and (b) if most of the words of a given sentence/snippet are more general than the words of the other sentence/snippet. As for

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

as we know, we are the first to propose an unsupervised language-independent asymmetric similarity measure in the context of Textual Entailment by Generality, although the approach from [4] is based on similar assumptions. As a consequence, it is likely that this methodology can be applied in heterogeneous text environment like the WWW.

This new proposal is exhaustively evaluated against the RTE-1 data set by testing different asymmetric association measures (Added Value, Braun-Blanket, Certainty Factor, Conviction, Gini Index, J-measure, Laplace and Conditional Probability) in combination with the $AISs$, with and without the identification of Multiword Units (MWU), which has proved to lead to improved results in [9]. In particular, we chose the RTE-1[2] as it is the only data set, which provides a suitable framework for the purpose of our research, by delivering three specific tasks i.e. Comparable Documents, Reading Compression and Paraphrase Acquisition. Some illustrative examples are presented below to better understand the tasks being tackled.

**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>

**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>

**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>

Finally, the evaluation shows promising results and evidences that the combination of the $AISs$ with the Added Value steadily improve results over other combinations. Moreover, the introduction of MWU identification shows different results for different tasks, which do not allow definitive conclusions.

## II. Related Works in RTE-1

Different approaches have been proposed to recognize Textual Entailment: from unsupervised language-independent methodologies [4] [5] [6] to deep linguistic analyses (an overview can be found in [3]). In this section, we will particularly mention the unsupervised language-independent approaches, which can be directly compared to our proposal, to some extent.

One of the most simple proposal is the one proposed by [5] who explore the BLEU algorithm. First, for several values of $n$, they calculate the percentage of $n$-grams from the text $T$, which appear in the hypothesis $H$. 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 to divide entailments from no entailments. A second more interesting work is proposed by [6], where the entailment data is treated as an aligned translation corpus. But, as the alignment scores alone were next to useless, they introduced a combination of string similarity metrics assembled by MITRE intended to measure translation quality. Finally, in order infer entailment, they used a K-NN classifier for K=5. The most interesting work is certainly the one described in [4] who propose a general probabilistic setting that formalizes the notion of Textual Entailment. 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, $\lambda$, which was set to 0.005 for best performance.

Although all three approaches show interesting properties, they all depend on tuned thresholds or fixed parameters, which can not reliably be reproduced. Moreover, some need training data, which may not be available. Our idea aims at (1) generalizing the hypothesis made by [4] by taking all pairs of words instead of just the one which maximizes the asymmetry between both sentences for each hypothesis word, (2) avoiding the definition of a "hard" threshold and (3) exhaustively studying asymmetry in language i.e. not just the conditional probability as done in [4] but many other asymmetric association measures.

## III. Asymmetry between Words

New trends have recently emerged with the study of asymmetric measures [8]. 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]. Here, we present eight asymmetric association measures that will be tested: Conditional Probability (Eq. 2), Added Value (Eq. 3), Braun-Blanket (Eq. 4), Certainty Factor (Eq. 5), Conviction (Eq. 6), Gini Index (Eq. 7), J-measure (Eq. 8) and Laplace (Eq. 9).

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

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

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

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

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

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

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

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

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

## IV. Asymmetry between Sentences

Although there are many asymmetric similarity measures between words, not many attributional similarity measures exist capable to assess whether a sentence/snippet is more specific/general than another one. As far as we know, the only exception is the measure proposed by [4], which shows lack of generality as mentioned in the previous section. To overcome this issue, we introduce the Simplified Asymmetric InfoSimba ($AISs$), 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 III.

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

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.

## V. Results and Discussion

In order to perform our evaluation, we first selected the Comparable Document (CD), Paraphrase Acquisition (PP) and Reading Comprehension (RC) data sets, as they are the one, which most suit to our final objective i.e. the labeling of ephemeral clusters with the web snippet, which best summarizes and subsumes all other web snippets in the cluster. All collections are balanced and the PP collection contains 50 examples, the CD collection, 150 examples and the RC collection, 140 examples. In Table I, we show the results of accuracy for all the data sets individually and the average accuracy.

Table I
Accuracy by Data Set Without SENTA.

| | CD | RC | PP | AVG |
|---|---|---|---|---|
| Glickman et al. [4] | **0.83** | 0.53 | 0.52 | **0.63** |
| Glickman et al. (keeping stop words) | 0.53 | 0.48 | 0.44 | 0.48 |
| Added Value (Equation 3) | 0.49 | 0.51 | 0.60 | 0.55 |
| J-measure (Equation 8) | 0.49 | 0.52 | **0.62** | 0.54 |
| Braun-Blanket (Equation 4) | 0.47 | **0.54** | **0.62** | 0.54 |
| Laplace (Equation 9) | 0.46 | 0.50 | 0.54 | 0.50 |
| Perez et al. [5] | 0.70 | 0.46 | 0.46 | 0.54 |
| Certainty Factor (Equation 5) | 0.47 | 0.51 | 0.52 | 0.50 |
| Conditional Probability (Equation 2) | 0.46 | 0.50 | 0.54 | 0.50 |
| Gini Index (Equation 7) | 0.47 | 0.46 | 0.40 | 0.44 |
| Conviction (Equation 6) | 0.49 | 0.50 | 0.48 | 0.49 |

On average, [4] shows the best results with 63% accuracy compared to the combination of the $AISs$ with the Added Value, which reaches 61%. However, when analyzing the results of [4] in more details, 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 [4] for the RC collection with the Braun-Blanket (54%) as well as for the PP collection with the J-measure and Braun-Blanket (62%). Moreover, when applying the methodology proposed by [4] keeping stop words, results drastically decrease and show second worst results on average after the Gini Index. Moreover, our best average results are obtained with the Added Value with 55% accuracy. This result is particularly interesting as it shows that the conditional probability alone may not be a good indicator to tackle specific entailments. Indeed, comparing two words based on the conditional probability does not guarantee a general to specific association as very frequent words provide high conditional probabilities and misjudge directed associations. The Added Value proposes a simple solution to this problem by subtracting the marginal probability (see Eq. 3).

It has been shown in similar experiments [9] that the informative-based similarity measures can gain in accuracy with the introduction of previous MWU identification. To confirm these results, we first processed each data set with the SENTA software [10] to identify their relevant MWUs. SENTA is a language-independent, threshold-free software, which runs keeping stop words. As such, we keep our original settings i.e. working with unmodified raw texts. The results are presented in Table II.

Table II
ACCURACY BY DATA SET WITH SENTA.

| | **CD** | **RC** | **PP** | **AVG** |
|---|---|---|---|---|
| Glickman et al. [4] | **0.83** | 0.53 | 0.52 | **0.63** |
| Glickman et al. (keeping stop words) | 0.54 | 0.52 | 0.44 | 0.50 |
| Added Value (Equation 3) | 0.53 | 0.49 | **0.62** | 0.55 |
| J-measure (Equation 8) | 0.41 | 0.51 | 0.56 | 0.50 |
| Braun-Blanket (Equation 4) | 0.48 | 0.45 | 0.58 | 0.54 |
| Laplace (Equation 9) | 0.51 | 0.46 | 0.54 | 0.50 |
| Perez et al. [5] | 0.70 | 0.46 | 0.46 | 0.54 |
| Certainty Factor (Equation 5) | 0.48 | 0.51 | 0.60 | 0.53 |
| Conditional Probability (Equation 2) | 0.51 | 0.46 | 0.54 | 0.51 |
| Gini Index (Equation 7) | 0.48 | **0.54** | 0.48 | 0.50 |
| Conviction (Equation 6) | 0.50 | 0.52 | 0.38 | 0.47 |

The overall results do not change with the integration of MWU as the best results on average are still obtained for the Added Value with the exact same accuracy i.e. 55%. However, a deeper look at the results show interesting issues. For the CD collection, almost all measures gain with the introduction of MWU, while for the RC collection, worst results are almost always found when compared to the results in Table I. The situation is mixed for the PP task. Moreover, we see that different measures provide best results for each task. The Added Value shows the best results of our methodology both for the CD and the PP tasks, which was not the case for the first experiment. Additionally, the Gini Index provides the best results for the RC collection. In particular, it seems that the measure that best benefits from the introduction of MWU is the Gini Index and the one which most looses is the J-measure. Although, these results are interesting they are not conclusive. For that purpose, we performed a new test, which consists in evaluating the results if we had chosen for each task the best results. In these conditions, we would reach 56% accuracy with the identification of MWU and 55% without. So, it seems that we may slightly benefit from the introduction of MWU.

## VI. CONCLUSIONS AND FUTURE WORKS

One possible solution to increase the expressiveness of cluster labels within the ephemeral paradigm is to provide each cluster with a small descriptive summary. Our idea is to select the "best" web snippet inside the cluster, which semantically subsumes all other ones. For that purpose, we proposed a first step towards this objective based on the detection of Textual Entailments by Generality. In particular, we proposed a new attributional similarity measure, called the asymmetric InfoSimba similarity measure ($AIS$), capable of assessing whether a sentence/snippet 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 improved results over [4] in a totally language-independent framework (i.e. keeping stop words). The next step of our study aims at proposing a graph-based framework to define which web snippet would summarize and subsume most of the other web snippets in a given cluster. For that purpose, we aim at building a graph where web snippets are vertices and directed edges between two vertices represent their entailment. A graph-based algorithm, such as the TextRank, would then order the web snippets by entailment relevancy.

## REFERENCES

[1] Buenaga, M. and Gachet, D. and Maña, M. and Villa, M., 2008. *Clustering and Summarizing Medical Documents to improve Mobile Retrieval*, Proceedings of the Workshop on-Mobile Information Retrieval associated with the 31st Annual International ACM SIGIR Conference, 38(2), 189-225.

[2] Bos, J. and Markert, K. 2005. *Recognising Textual Entailment with Logical Inference*. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005), Association for Computational Linguistics, Morristown, NJ, USA, 628-635.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] 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.

[9] 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.

[10] Dias, G. 2002. *Extraction Automatique d'Associations Lexicales à Partir de Corpora*, PhD Thesis, Univeristy of Orléans and New University of Lisbon.