A Simple Ontology Based on Text Entailment Directional Relationship

Andreea-Diana MIHIŞ
Cluj-Napoca, România
mihis@cs.ubbcluj.ro

Natural language processing techniques are becoming more and more accurate and so, reliable to be used in other processes. For instance, natural language tools where used in software engineering for test generation [5], for a requirements ontology generation [6], and in [4], is proposed a tool based on text similarities capable to improve the specification document quality, by checking for inconsistencies and ambiguities.
***Keywords**: ontology, Text Entailment, natural language, document quality.*

Ontology's definition has a philosophy definition, meaning the science and the study of being (ontos – existence, logos – science), [3]. In general, any knowledge statement is based on the existence of an ontology, which means that it is starting from the basic categories of entities and relations identification of the target domain.

Although the scope of an ontology creation process is to create an formal ontology, a tool supported one, there are informal ontologies too. They use natural language for the definition of the entities and relations, and they can be obtained automatic or semi-automatic using a natural language processing tool. In 1993 Bateman, in [1] notice the potential of using natural language processing mechanisms in the creation of a informal or formal ontology. In 1996, Mike Uschold published the results of his work in converting a informal ontology written in natural language into Ontolingua [11].

Meanwhile, the natural language techniques improved, so, a informal ontology is easier to be used or transformed into a formal one, or even created, such is a temporal ontology (the relation between entities is represented by time), or, why not, even a text entailment relationship based ontology.

The text entailment relation between two texts: T (the text) and H (the hypothesis) represents a fundamental phenomenon of natural language. It is denoted by $T \rightarrow H$ and understands that the meaning of H can be inferred from the meaning of T. The recognition of textual entailment is one of the most complex tasks in natural language processing and the progress on this task is key to many applications such as question answering, information extraction, information retrieval, text summarization, and others [10].

The Text Entailment Directional Relationship

The importance and the utility of this relation is so great, that in the last years, several competitions named Recognizing Text Entailment (RTE) competitions where organized [9]. All have the purpose to decide if a text T entails another text H (the hypothesis), relation logically denoted by $T \rightarrow H$, but, unlike first two competitions, for which the result of an entailment relation was true or false, for the last two competitions, inclusive this year competition, the results are „true", „possible" and „improbable". The actuality and the importance of this Natural Language Processing relationship are given by the interest of the researchers who organized and participate to the challenges.

In this paper I use a directional Text Entailment recognizing technique, presented in [10]. This technique was tested on the first RTE-1 competition dataset and the precision of this technique was 57.62%, comparable to the precision obtained by Glickman [2], 58.5%, the winner of the RTE-1 Challenge. This technique uses three cosine measures. They consider the words of $T = t_1, t_2, \ldots, t_m$ and of $H = h_1, h_2, \ldots, h_n$. These words are in fact the distinct radix words, where for the radix identification was used a dictionary.

The two vectors for calculating cos_T (*T*,*H*) are: $\vec{T}$ = (1,1,…,1) (a *m*-dimensional vector) and $\vec{H}$, where $\vec{H}_i$ = 1, if t_i is a word in sentence *H* and $\vec{H}_i$ = 0 otherwise.

The two vectors for calculating cos_H(*T*,*H*) are: $\vec{H}$ = (1,1,…,1) (a *n*-dimensional vector) and $\vec{T}_i$ = 1, if h_i is a word in sentence *T* and $\vec{T}_i$ = 0 otherwise.

For $cos_{H\cup T}$ (*T*,*H*) the first vector is obtained from the words of *T* contained in $T\cup H$ and the second, from the words of *H* contained in $T\cup H$.

Denoting by *c* the number of common words of *T* and *H*, the three measures are: $cos_T(T,H)=\sqrt{\frac{c}{m}}$, $cos_H(T,H)=\sqrt{\frac{c}{n}}$ and $cos_{H\cup T}(T,H)=\sqrt{\frac{4c^2}{(n+c)(m+c)}}$.

Relations between them are: $\cos_H(T,H) \geq \cos_{H\cup T}(T,H) \geq \cos_T(T,H)$, considering $m \geq n \geq c$. Namely, for 94% from the dataset of pairs relation $\cos_H(T,H) \geq \cos_T(T,H)$ holds, for 97% relation $\cos_H(T,H) \geq \cos_{H\cup T}(T,H)$ holds and for 76% relation $\cos_H(T,H) \geq \cos_{H\cup T}(T,H) \geq \cos_T(T,H)$ holds. The reason is that $\cos_{H\cup T}(T,H) \geq \cos_T(T,H)$ only if $c \geq m/3$ and this is fulfilled only for 77% of total set of pairs $T\cup H$.

To accomplish the condition: *T entails H iff H is not informative in respect to T* [7], the similarity between *T* and *H* calculated with respect to *T* and to $H \cup T$ must be very closed. Analogously, the similarity between *T* and *H* calculated in respect to *H* and to $H \cup T$ must be very closed. Also, all these three similarities must be bigger than an appropriate threshold. The conditions imposed are:

$|\cos_{H\cup T}(T,H) - \cos_T(T,H)| \leq \tau_1$

$|\cos_H(T,H) - \cos_{H\cup T}(T,H)| \leq \tau_2$

$\max\{\cos_T(T,H), \cos_H(T,H), \cos_{H\cup T}(T,H)\} \geq \tau_3$

The threshold founded by a learning method was: τ_1=0.095, τ_2=0.15 and τ_3=0.7.

Because τ_1 and τ_2 are not equal, this text entailment relation is a directional one.

Statistics for the accuracy and average precision obtained by tasks, are given in the next figure:

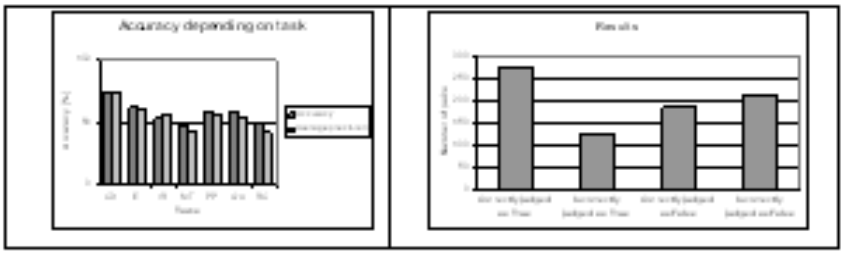


Fig.1. The Cosine Text Entailment technique evaluation

Namely, for CD the accuracy is 73.64864865% (average precision 74.71911819%), for IE is 61.66666667% (61.08168133%), for IR is 52.80898876% (55.38470939%), for MT is 47.5% (42.08519059%), for PP is 58.82352941% (56.47843628%), for QA is 58.46153846% (54.41986024%), and for RC is 48.20143885% (41.65598715%). Let remark that the best score for CD task is almost a permanent feature for the systems participating at RTE-1.

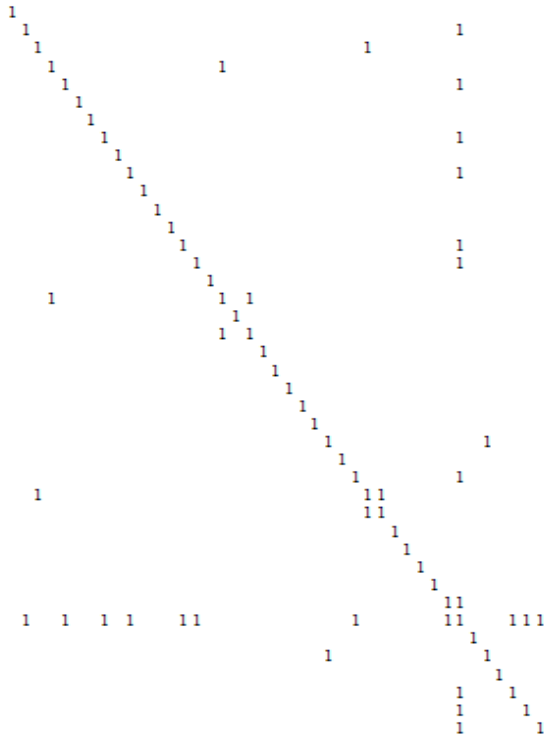
The accuracy for TRUE pairs is 68.92230576% and for FALSE pairs is 46.36591479%. The global accuracy is 57.62%.

The text entailment method tool use WordNet's on-line dictionary. This method and tool can easily be used for another language, simply by changing the dictionary.

Text Entailment and Ontology

Because Text Entailment by cosine technique is a directional relation between texts, it can be used to identify the relations between entities written in natural language in an informal

ontology.
Starting from a natural language text, if for all the pairs of sentences, the inter-sentence's text entailment relations are considered, a sentence directional relationship graph will be obtained. This graph can be represented in a matrix form, such as the following matrix, obtained from the newspaper text used in [8], text named in the following Hirst text.



Although the text entailment relation is a directional one, in some cases it will be a bidirectional relation. Such a case is the case of self-entailment. For the Hirst text, all the entailment relationships are bidirectional ones. It can be noticed that the 31st sentence is the most important one.
From this graph, a simple, informal, natural language ontology of the tested text can be obtained. In this simple ontology, the entities are the text's sentences, and the relations are text entailment relations. This ontology is not a tree-type one, but a tree ontology can be obtained simply by starting from one sentence and stopping when the text entailment relation bind a low level sentence to a upper level sentence one. This tree ontology can start form a text's sentence, or from a sentence which didn't belong to the text.
For instance, starting from the 19th sentence of the text, this sentence entails sentence 17 that entails sentence 4. The obtained tree is a linear one: 19 – 17 – 4.
Even if in this example the sentence's text entailment was used, the text entailment relation can be used between different texts. In this way, not a sentence's text entailment graph will be obtained, but a text's text entailment graph. If the texts represent entities definitions, graph obtained by text entailment will represent the relationship's graph.

Conclusions

In the paper, was presented the utility of text entailment relationship in the creation of an informal ontology. This ontology can be combined with other natural language ontology, such as a similarity based ontology or a word count based ontology, in order to obtain the best results.
The advantage of this approach is that it can be applied on natural language texts, and, even if the resulting ontology is a formal one, it can represent the starting point for the identification of a formal ontology. The text entailment directional identification method is a language-independent one, and so, it can be used for different language texts.

References

[1] Bateman, J.A., *Ontology Construction and Natural Language*, Proceeding of the International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation, editors N. Guarino and R. Poli, Padova, March 1993, pp. 83-93
[2] Glickman, O., Dragan, I. şi Koppel, M., *Web Based Probabilistic Textual Entailment*, Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, 2005
[3] Gruber, T. 1996. „What is an Ontology", http://www.kr.org/top/definitions.html
[4] Harksoo, K., Youngjoong, K., Sooyoung, P., Jungyun, S., *Informal Requirements Analysis Supporting System for Human Engineer*, Proceedings of Conference on IEEE-SMC99, Vol. 3, 1999, pp. 1013-1018

[5] Lutsky, P., *Artificial Intelligence in Engineering*, Volume 14, Issue 1, January 2000, pp. 63-69

[6] Kof, L., *Natural Language Processing: Mature Enough for Requirements Documents Analysis*, Lecture Notes in Computer Science, 2005, pp. 91-102

[7] Monz, C. şi M. de Rijke, *Light-Weight Entailment Checking for Computational Semantics*, Proceedings IcoS-3, editors Blackburn P. and Kohlhase M., 2001

[8] Morris, J., Hirst, G., *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*, Computational Linguistics, Vol. 17, numbar 1, 1991, pag. 21-48

[9] http://www.pascalnetwork.org/Challenges/RTE/

[10] Tătar, D., Şerban, G., Mihiş, A.D., Mihalcea, R., *Textual Entailment as a Directional Relation*, CALP2007, INCOMA Ltd., editors Orăştean Constantin and all, pag. 53-58

[11] Uschold, M., *Converting an Informal Ontology into Ortolingua: Some Experiences*, ECAI 96, Proceedings of the Workshop on Ontological Engineering, Vol. 192, Budapest, March 1996, pages 1-17