UAIC Participation at RTE4

Adrian Iftene

„Al. I. Cuza“ University, Faculty of Computer Science, Iasi, Romania

adiftene@info.uaic.ro

Abstract

Textual entailment recognition is the task of deciding, when given two text fragments, whether the meaning of one text is entailed from the other text. This year, at our second participation in the RTE competition, we improve the system built for the RTE3 competition. The main idea of our system is to map every word from hypothesis to one or more words from the text. For that, we transform the hypothesis making use of extensive semantic knowledge from sources like DIRT, WordNet, VerbOcean, Wikipedia and Acronyms database. After the mapping process, we associate a local fitness value to every word from hypothesis, which is used to calculate a global fitness value for current fragments of text. The global fitness value is decreased in cases in which a word from hypothesis cannot be map to one word from the text or when we have different forms of negations for mapped verbs. In the end, using thresholds identified in the training step for global fitness values, we decide for every pair from test data if we have entailment or not.

1. Introduction

The RTE4[1] track at TAC 2008 continues the previous RTE Challenges that have aimed to focus research and evaluation on underlying semantic inference task. The goal of the RTE Track is to develop systems that recognize when one piece of text entails another. This year, similar to the track[2] piloted in RTE-3 from 2007, the 3-way classification was included. The goal of making a three-way decision of “Entailment”, “Contradiction” and “Unknown” is to drive systems to make more precise informational distinctions; a hypothesis being unknown on the basis of a text should be distinguished from a hypothesis being shown false/contradicted by a text.

The system used this year represents an improvement version of the system from RTE3 (Iftene and Balahur-Dobrescu, 2007). Additionally, we added new rules and used new semantic resources with the aim of better identifying the contradiction cases. Figure 1 shows the actual system (with gray are the new added components):

[1] RTE-4: http://www.nist.gov/tac/tracks/2008/rte/

[2] RTE-3 pilot: http://nlp.stanford.edu/RTE3-pilot/

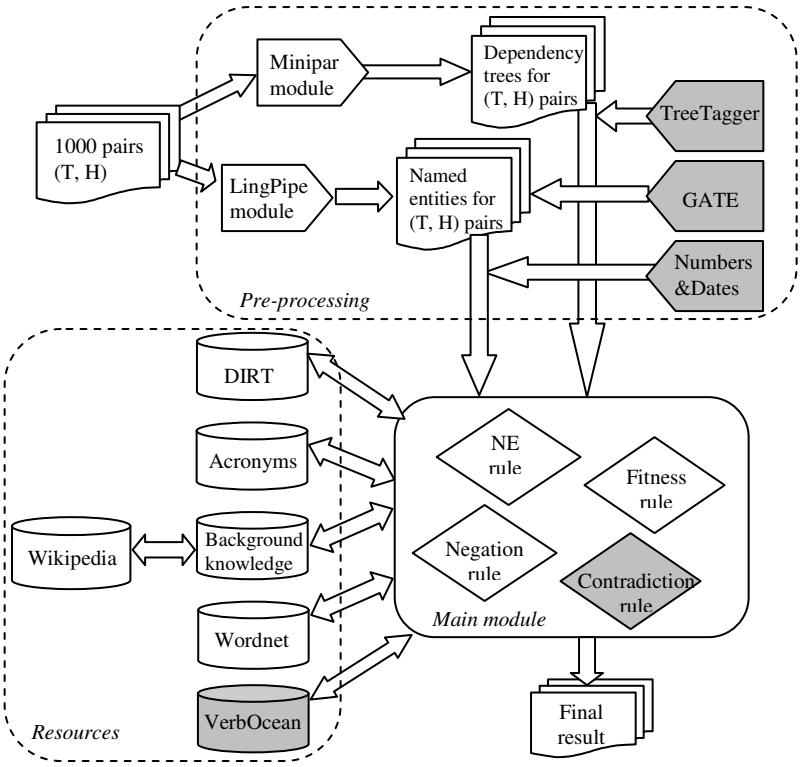

Figure 1: System architecture

Next chapters present the main components of the system and in the end, obtained results from this year competition are presented. We will insist on new added components and on new used resources.

2. Pre-Processing

Initially, the text and the hypothesis are transformed with MINIPAR (Lin, 1998) tool into dependency trees: dependency tree associated to the text (called *text tree*) and dependency tree associated to the hypothesis (called *hypothesis tree*). Because in some cases the part-of-speech (POS) identified by MINIPAR is wrong, we use a TreeTagger tool[3] that correctly identifies the words POS and replaces the wrong POS identified by MINIPAR. This step is very important, especially for verbs, because our algorithm starts from verbs mapping and all the next steps depend on it (Iftene, 2008).

In order to identify name entities, all text-hypothesis pairs are sent to the LingPipe[4] module. In order to improve the results obtained in RTE3, in the case of Named Entities of type PERSON, we additionally used GATE (Cunningham et al., 2001), which contains finer-grained classes of entities. In a manner that is similar to the approach presented in (Tatu and Moldovan, 2007) we distinguish between the cases in which the *family name* or the *first name* found in the hypothesis are missing from the text. In addition to using LingPipe and GATE, we build a set of patterns

[3] TreeTagger: http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php
[4] LingPipe: http://www.alias-i.com/lingpipe/

with aim to identify numbers and calendar dates. In the RTE3 system we missed some situations, but now we add specific patterns for the identification of *percentages*, *measure values*, *time periods*, *order numerals*, etc. (Iftene, 2008).
In both trees for every node (which represents one word from initial sentence) we have the following information: the word lemma, the word part-of-speech (verb, noun, adjective), a flag *isNameEntity* (*true* if the current word is a name entity and *false* otherwise), other information. For nodes from the hypothesis tree we have some additional information regarding corresponding node from text tree, the current node to which is mapped and its local fitness.

3. Main Module

The main objective is to map every node from the hypothesis tree to one node from the text tree, similar to our approach from RTE3 (Iftene and Balahur-Dobrescu, 2007). It is possible that one node from the hypothesis tree to be map to several nodes from the text tree. For this reason, we will consider, in both dependency trees, triplets in the form (node lemma, father lemma, edge label) named *entity*. From this moment on, we will try to map entities from the hypothesis tree to entities from the text tree.
The mapping between entities can be done in two ways: *directly* (when entities from hypothesis tree exist in the text tree) or *indirectly* (when entities from text tree or hypothesis tree cannot be mapped directly and support transformations using external resources in order to do the mapping). Using this type of mapping between an entity from hypothesis and an entity from text, we calculate a *local fitness* value which indicates the appropriateness between entities. Based on local fitness, we build an *extended local fitness* and in the end, using all partial values, we calculate a normalized value that represents the *global fitness*. When an entity from the hypothesis tree can be mapped to more entities from the text tree, we select the mapping which increases the global fitness with the highest value.
Some of the following steps help our program in identifying the final answer for every pair from our test data. Accordingly, with an applied tool or with used resources we increase or decrease the global fitness score, and in the end we will have the final answer. The “No entailment” cases are represented by pairs of text and hypothesis for which we have the value of global fitness under a threshold identified on training data, and the “Entailment” cases are represented by pairs for which we have global fitness over the same threshold. In order to differentiate contradictions and unknown cases, we considered another threshold, also identified on training data. Since penalties for *contradiction cases* are higher than penalties for *unknown cases*, the order is as follows: the lower level for global fitness, the contradiction cases, first threshold, unknown cases, the second threshold, the entailment cases, and the highest level for global fitness. (See in Figure 2).

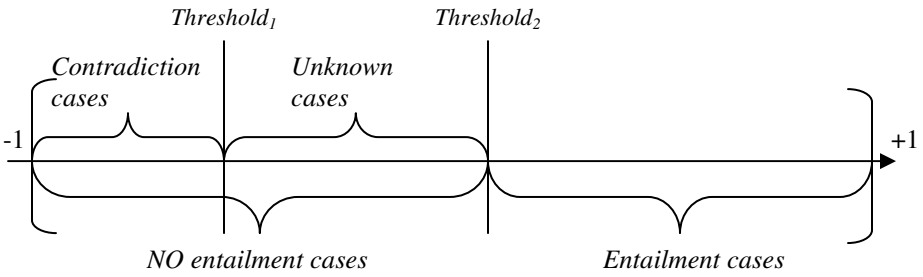

Figure 2: Final answers consideration using Global fitness values

Let's see for every type of possible answer, what the rules that promote it are.

3.1. Entailment Cases

3.1.1. Basic Positive Rules

Of course, in this case any type of mapping will increase the global score and, in the end, will increase the probability to have the final answer "Entailment". Let's see how we calculate the local fitness in these situations. For every node from the hypothesis tree which can be mapped directly to a node from the text tree, we will consider the local fitness value to be 1 (which represents the maximum value).

When it is not possible to do a direct mapping between a hypothesis node and a node from the text, we will try to use the external resources and to transform the hypothesis node in an equivalent one. Thus, for *verbs* we use the DIRT resource (Lin, 1998) and transform the hypothesis tree into an equivalent one, with the same nodes except the verb. This is the case of pair 84 where in text we have "*Barack Obama has declared …*" and in hypothesis we have "*Obama claims Democratic victory*" and we found in DIRT a similarity relation between verbs *declare* and *claim* with score 0.290608. After using this resource, the hypothesis has changed in "*Obama declares Democratic victory*" and in this form it is easier to compare the text and hypothesis and in the end the value of the global fitness score is increased.

If the word is *named entity*, we try to use an acronyms' database[5] or obtain information related to it from the background knowledge (Iftene and Balahur, 2008). As an example for acronyms we can indicate pair 528 where in text we have *EU* and in hypothesis we have *European Union*. Examples in which we use our module which adds new elements from English Wikipedia to the background knowledge are pairs 51 (relation between *Jewish* and *Jew*), 104 (between *Buenos Aires* and *Argentina*), 184 (between *Ireland* and *Irish*), 216 (between *UK* and *British*), 280 (between *16* and *sixteen*), 528 (between *Portugal* and *Portuguese*), etc.

For *nouns* and *adjectives* we use WordNet (Fellbaum, 1998) and a part of the relations from eXtended WordNet[6] to look up synonyms and then we try to map them to nodes from the text tree.

[5] Acronyms: http://www.acronym-guide.com
[6] eXtended WordNet: http://xwn.hlt.utdallas.edu/

We only took relations with high scores of similarity from eXtended WordNet. Examples of synonymy relation from WordNet are pairs 57 (between *nuclear* and *atomic*), 92 (between *trouble* and *problem*), 114 (between *talk* and *discussion*), etc.
For every transformation with DIRT or WordNet, we will consider the similarity value indicated by these resources for local fitness. When we use the acronyms database or background knowledge we consider the local fitness 1.

3.1.2. Positive Rules for Numbers

In the test data from this year we observed special situations for name entities of numeric type. Even if we don't have the same numbers in text and hypothesis, we have some quantification words before or after numbers change their meaning and in the end we have equivalent relations between them. For example, at pair 304 we have in text "*at least 80 percent*" which is equivalent with "*more than 70 percent*" from hypothesis. With our initial rule we compare only numbers (80 and 70 in this case) and the final answer depends on the fact that these numbers are different. With the new rule, we create intervals for both expressions (percents over 80 for text and percents over 70 for hypothesis) and because these intervals have common elements, the final answer is different in this case. The list of quantifications words contains expressions like "more than", "less than", or words like "over", "under" and in these cases we consider intervals instead of considering only numbers.
Another situation is the case at pair 331 where we have in text "*killing all 109 people on board and four workers on the ground*" and in the hypothesis we have "*killed 113 people*". How we can see if we add "109" and "four" from text we obtain "113" that represent the value from hypothesis. For the same reason as above, we add a new rule which takes into consideration an additional number belonging to one sentence (text or hypothesis) which is obtained as sum of consecutive numbers separated by the word "and" or commas.

3.2. No Entailment Cases

3.2.1. Basic Negative Rules

If after all checks are made we cannot map one node from the hypothesis tree, we insert some penalty in the value of the node's local fitness. Also, because the stop words from the hypothesis ("the", "an", "a", "at", "to", "of", "in", "on", "by", etc.) artificially increase the value of global fitness, we don't take them into consideration in the final global fitness.

3.2.2. Negation Rules

For every verb from the hypothesis we consider a Boolean value which indicates whether the verb has a negation or not. For that, we check inside its tree on its descending branches to see whether one or more of the following words are to be found: "not", "never", "may", "might",

"cannot", etc. For each of these words we successively negate the initial truth value of the verb, which by default is "false". The final value depends on the number of such words.
Another rule was built for the particle "to" when it precedes a verb. In this case, the sense of the infinitive is strongly influenced by the active verb, adverb or noun before the particle "to", as follows: if it is being preceded by a verb like "believe", "glad", "claim" or their synonyms, or adjective like "necessary", "compulsory", "free" or their synonyms or noun like "attempt", "trial" and their synonyms, the meaning of the verb in infinitive form is stressed upon and becomes "certain". For all other cases, the "to" particle diminishes the certainty of the action expressed in the infinitive-form verb.
We will see below more exactly how these rules were split into rules for contradiction cases and rules for unknown cases.

3.2.3. Contradiction Cases

From negation rules, we consider *contradictions cases*: when verbs are negated with words like "never", "not", "no", "cannot", "unsuccessfully", "false" etc. (Iftene, 2008). This case is encored at pair 660 where in text we have "*Aquacell Water, Inc announced today that it has* ***not received*** …" and in hypothesis we have "*Aquacell Water* ***receives*** …".
Also, we consider cases when before particle "to" we have words like "refuse", "deny", "ignore", "plan", "intend", "proposal", "able", etc. These situations appear for example at pair 54 where the text is "***Plans to detain*** *terrorist suspects for up to 42 days without charge* …" and the hypothesis is "*Police* ***can detain*** *terror suspects for 42 days without charge.*", and at pair 354 where in text we have "… *Shin was sacked from the school on June 20 after* ***refusing to resign*** *from his post as director of KBS*." and the hypothesis is "*Shin Tae-seop* ***resigned*** *from his post at Dong-eui University*.".
An important situation in contradiction identification is determined by identification of antonymy relation between words from text and hypothesis. In order to identify this relation we use [*opposite-of*] relation from VerbOcean resource (Chklovski and Pantel, 2004) and antonymy relation from WordNet. Examples of opposite relations from VerbOcean are for pairs 8 (between *increase* and its opposite *decrease*), 885 (between *pay* and its opposite *sell*). Examples of antonymy relation from WordNet are for pairs 28 (between *low* and *increase*) and for 48 (between *leave* and *stay*).
In order to catch more situations, we consider a combination between synonyms from WordNet and antonymy relation from WordNet or opposite relation from VerbOcean. So, using only the WordNet, for words from hypothesis which cannot be mapped to words from text using synonymy relation or antonymy relation, we consider the set of antonyms for their synonyms and we check if something from this new set can me mapped to the text. This is the case of pair 302, where we have in hypothesis the word *separate* which is synonym with the word *distinct*. The antonym of the word *distinct* is the word *same* which appear in the text. Thus, between words *separate* from hypothesis and word *same* from text we have an antonymy relation, even if we

don't have this explicit relation in WordNet. Similarly, we consider corresponding sets using the synonyms from WordNet and opposite relation from VerbOcean.
In some situations, similarity relation from DIRT is antonymy relation, and for this reason we do an extra verification of DIRT relations to see if we have antonymy in WordNet or in VerbOcean. If we have an antonymy relation, we will change the local fitness of the node with a negative value. For example, initially at pair 167 we have, using DIRT, a relation between *convict* and *acquit* with score 0.302455, but because we found in WordNet that *convict* and *acquit* are antonyms, we change the local fitness score of hypothesis verb and we insert a penalty.
For all identified *contradiction cases*, since we consider the penalties with the highest values, in the end, the final answer for the considered pairs will be "*Contradiction*".

3.2.4. Unknown Cases

From the negation rules, we consider *unknown cases*: when words are "may", "can", "should", "could", "must", "might", "infrequent", "rather", "probably", etc. (Iftene, 2008). In these cases, inserted penalties are not decisive in final answer establishing, which is obtained only after the calculation of global fitness. At pair 198 we have in text "… *could also be linked to* …" and in hypothesis we have "… *is linked to* …".
Related to the particle "to" we will consider the cases which are not included in contradiction cases. So at pair 391 with text "*It* ***is hard to like*** *Will Carling* …" and hypothesis "*Nobody* ***likes*** *Will Carling*" we insert a penalty.
If at the cases presented above we only insert penalties and only in the end, after the calculation of the global fitness, will we know the final result for a pair, things are different regarding *named entities* with problems. In the event that even after using acronyms database and background knowledge we cannot map the *named entity* from the hypothesis to a named entity from the text tree, we decide that the final result for the current pair: "*Unknown*". This case is encored at pair 454 from bellow, in that we identify the named entity *Russia* in hypothesis without a corresponding value in the text.

> T: *In 1977 to wide media fanfare, Polanski was charged with a host of sexual crimes for his involvement with a 13-year-old girl. He was subsequently convicted of unlawful intercourse with a minor, but fled the country in 1978 before final sentencing.*
>
> H: *Polanski fled from the U.S. to Russia in 1978.*

For the numbers from text and from hypothesis, when it is possible, we also keep their unit measure. The rule presented above for named entities is also applied in cases in which we have the same numbers in text and in hypothesis, but we have different unit measures for them. This is the case for pair 441 where we have *11 troops* in the hypothesis and *11 September* in the text:

> T: *Britain deployed troops to Afghanistan shortly after the attacks of 11 September, 2001. Few then thought that British forces would still be in Afghanistan in far larger numbers seven years on, nor that they would be involved in some of the fiercest fighting British forces have seen in decades, as part of Nato's International Security and Assistance Force (ISAF).*

H: *Britain has 11 troops that take part in Nato's International Security and Assistance Force.*

An exception from the named entity rule presented above is the case when the type of name entity is *first name*. In this case we only insert a penalty in the global fitness. This is the case of pair 122 from below where we have *Gordon Brown* in the hypothesis and we only have *Mr Brown* in the text:

T: *Mr Brown's courage and determination are not in doubt: he soaks up punishment as if he believes it is good for him. But week after week he gives no sign that he knows how to seize the initiative and dictate the course of the fight.*

H: *Gordon Brown is the UK Prime Minister.*

4. Results

The distributions of our answers in a 3-way task are presented below:

Answer Type	**# of answers in Gold**	**# of correct answers given by our system**	**Total # of answers given by our system**	**Precision**	**Recall**	**F-measure**
Entailment	500	466	712	65.45%	93.20%	76.90%
Contradiction	150	69	85	81.18%	46.00%	58.72%
Unknown	350	150	203	73.89%	42.86%	54.25%
Total	**1000**	**685**	**1000**	**68.50%**		

Table 1: Results in RTE4 on 3-way task

As we can see, the highest precision is for *Contradiction* case: 81.18 % and the lowest precision is for *Entailment* case: 65.45%. Also, we can see that the highest recall and F-measure are obtained for *Entailment* case 93.2% and 76.9 % and the lowest are for *Unknown* case, 42.86% and 54.25 %. The meaning of these results is that our system offers very many *Entailment answers* and catches almost all possible *Entailment* cases; also the system offers a lower number of *Contradiction* and *Unknown* answers, but almost all are correct.
For the 2-way task, the distribution is presented in table bellow:

Answer Type	**# of answers in Gold**	**# of correct answers given by our system**	**Total # of answers given by our system**	**Precision**	**Recall**	**F-measure**
Yes	500	466	712	65.45%	93.20%	76.90%
No	500	255	288	88.54%	51.00%	64.72%
Total	**1000**	**721**	**1000**	**72.10%**		

Table 2: Results in RTE4 on 2-way task

The results are similar to results from the 3-way task and we notice the very high precision for No cases (88.54%), where from 288 answers offered by our system 255 are correct. The meaning of the difference between global precision from 2-way task and 3-way task is that only in 36 cases from 255 cases we don't distinguish correctly between *Contradiction* and *Unknown* cases.
In comparison with results from RTE3 we can see that accordingly to the provenience of testing data we have significant improvements on IR, SUM and IE tasks, but on QA task, where we got the best results in RTE3, we have the lowest result in RTE4.

Provenience of testing data	**RTE3**	**RTE4**
IR	69.00 %	82.00 %
QA	87.00 %	63.00 %
SUM	63.50 %	78.00 %
IE	57.00 %	64.33 %
Total	**69.13 %**	**72.10 %**

Table 3: Comparison between results between RTE3 and RTE4

5. Conclusions

In this paper, the system used this year in RTE4 competition is presented. The main part consists of the presentation of ways in which we use the new tools (GATE, TreeTagger), new resources (VerbOcean) and new rules (Contradiction, Named Entities).
With the new changes presented, the system precision is improved, and the system is more oriented to an *Entailment-Contradiction-Unknown* system, because the rules for contradictions are more clearly specified. The obtained results from RTE-4 are better than the results from RTE3 participation: 72.1 % on two-way task (with 3 % better than in RTE3) and 68.5 % on three-way task (with 14.5 % better than in RTE3).
The main problems are related to cases in which text and hypothesis are very similar and contain the same set of words, but we have a different order for words which have different semantic roles in text and in hypothesis. The solution for this problem is to use a special tool that identifies semantic roles for words and to insert new rules for cases in which the same word has different roles in text and in hypothesis.

Acknowledgments

The author thanks the members of the NLP group in Iasi for their help and support at different stages of the system development. The work on this project is partially financed by the SIR-RESDEC, PNCDI II project and by Siemens VDO Iaşi.

References

Chklovski, T., Pantel, P. 2004. *VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations*. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. 2001. *GATE: an architecture for development of robust HLT applications.* In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2001, 168--175, Association for Computational Linguistics, Morristown, NJ, USA

Fellbaum, C. 1998. *Wordnet: An electronic lexical database*. MIT Press, Cambridge, Mass.

Iftene, A. 2008. *Building a Textual Entailment System for the RTE3 Competition. Application to a QA System*. In proceedings of 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2008). September 26-29, Timişoara, România.

Iftene, A., Balahur-Dobrescu, A. 2007. *Hypothesis transformation and semantic variability rules used in recognizing textual entailment*. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 125–130, Prague, June 2007. Association for Computational Linguistics.

Iftene, A., Balahur-Dobrescu, A. 2008. *Named Entity Relation Mining Using Wikipedia*. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 28-30 May, Marrakech, Morocco.

Lin, D. 1998. *Dependency-based Evaluation of MINIPAR*. In Workshop on the Evaluation of Parsing Systems, Granada, Spain, May, 1998.

Lin, D., Pantel, P. 2001. *DIRT - Discovery of Inference Rules from Text*. In Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-01). pp. 323-328. San Francisco, CA.

Tatu, M., Moldovan, D. 2007. *Cogex at RTE 3*. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 22–27, Prague, June 2007. Association for Computational Linguistics.