UAIC Participation at RTE5

Adrian Iftene, Mihai-Alex Moruz

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

{adiftene, mmoruz}@info.uaic.ro

Abstract

Textual entailment recognition is the task of deciding, given two text fragments, whether the meaning of one text can be deduced from the other. This year, at our third participation in the RTE competition, we improved the system built for the RTE4 competition.

Main Task: The main idea of our system is to map every word in the hypothesis to one or more words in the text. For that, we transform the hypothesis, using extensive semantic knowledge from sources like DIRT, WordNet, VerbOcean, Wikipedia and the Acronym database. The main improvement this year was related to the pre-processing part. Last year we observed how this part can improve the quality of the output for the tools used (LingPipe and Minipar). Because this year the texts were obtained from a variety of sources and were not edited from their source documents, we focused on this part. Thus, we identify and eliminate special characters that occur frequently on web pages. This choice is based on the fact that “with or without these characters the meaning of the text is the same, but the quality of the tools output is improved. Additionally, we process the LingPipe output with GATE in order to identify some named entities categories unidentified by LingPipe such as nationality, language, and job. One of the better components of last year’s system, the one responsible with the solving of contradiction cases, has not functioned properly this year. Also, cases in which the texts were very long and hypothesis were very short, but for which most of the words in the hypothesis were found in the text, were not treated properly by our system, because we did not use proper differences that come from semantic role labeling.

Pilot Task: Regarding the new pilot task introduced this year, we used Lucene in order to index documents in which we must identify sentences that entail a given hypothesis. On this index we performed searches using the initial hypotheses, and after filtering the results offered by Lucene, we applied our RTE system.

1. Introduction

The RTE5[1] track at TAC 2009 continues the previous RTE Challenges that have aimed to focus research and evaluation on underlying semantic inference task. Given two text fragments called 'Text' and 'Hypothesis', Textual Entailment Recognition is the task of determining whether the meaning of the Hypothesis is entailed from the Text. Since its inception in 2004, the RTE Challenges have promoted research in textual entailment recognition as a generic task that captures major semantic inference needs across many natural language processing applications. This year, the task was similar to RTE4 with two relevant changes: 1) the average length of the Texts was higher, and 2) texts come from a variety of sources, without additional processing from their source documents.

[1] http://www.nist.gov/tac/tracks/2009/rte/

The system used this year represents an improvement version of the previous systems from RTE3 (Iftene and Balahur-Dobrescu, 2007) and from RTE4 (Iftene, 2008). Additionally, we added new modules and used new semantic resources with the aim to deal with new changes from RTE track and also with aim of better identifying the unknown cases. Figure 1 shows the actual system (with gray are the new added components):

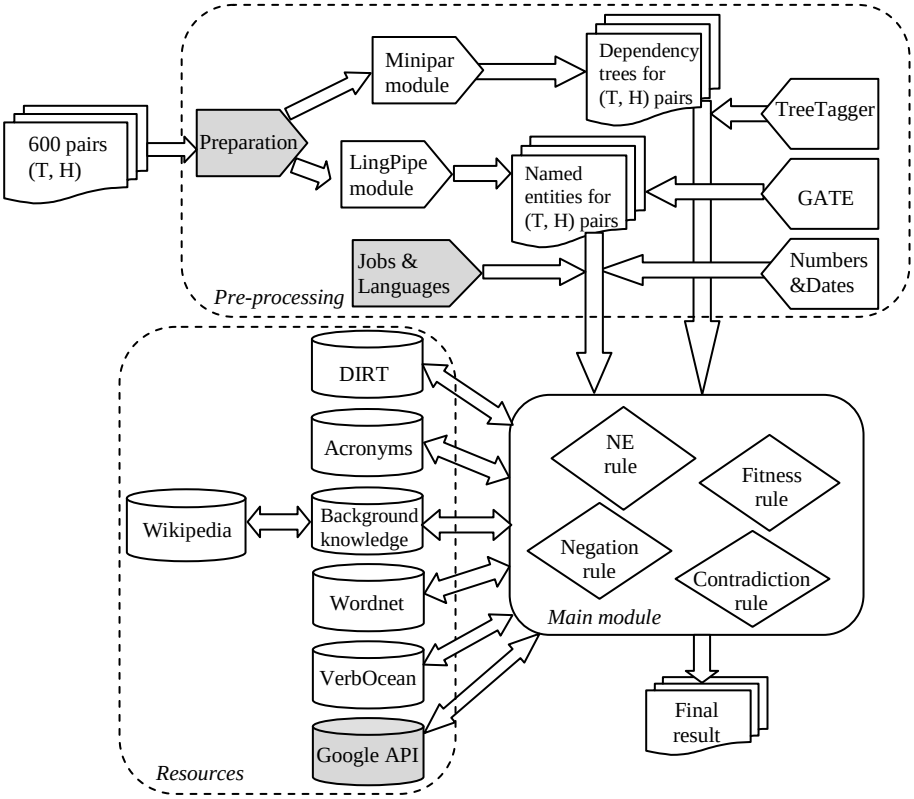

Figure 1: RTE5 System architecture

2. Pre-Processing

Preparation: In order to improve the quality of the preprocessing output, some additional steps are performing (Iftene, 2009). Thus, in all test data we replace "hasn't" with "has not", "isn't" with "is not", "couldn't" with "could not", etc. The meaning of the text remains the same after transformation, but the MINIPAR output is better for this new text. Also, before sending the text to LingPipe, we replace some punctuation signs like quotation marks "", brackets (), [], {}, commas, etc. with the initial sign between spaces. Again, the meaning of the text is the same, but the LingPipe output is better processed further after this transformation.

MINIPAR: after preparation step, the text and the hypothesis are parsed with MINIPAR (Lin, 1998). In cases in which MINIPAR doesn't identify any verb in the processed sentence, we use the TreeTagger tool[2] that identifies, with a higher degree of precision, the Part-Of-Speech (POS) and replaces the incorrect 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.

[2] http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php

LingPipe: In parallel, the result obtained after preparation was sent to be processed by LingPipe[3], in order to identify named entities. In order to improve the results obtained in RTE4 (Iftene, 2008), in the case of Named Entities of type JOB and LANGUAGE, we additionally used GATE (Cunningham et al., 2001), which contains finer-grained classes of entities.

3. Main Module

The main objective is to map every node from the hypothesis tree to one node from the text tree, in a similar manner as described in (Iftene, 2008). The mapping between entities can be done either *directly* (when entities from hypothesis tree exist in the text tree) or *indirectly* (when entities cannot be mapped directly and require transformations using external). Using this type of mapping, we calculate a *local fitness* value which indicates the similarity between entities of the text and the hypothesis. Using the local fitness values, we build an *extended local fitness* and then, using all partial values, we calculate a normalized value that represents the *global fitness*. When an entity from the hypothesis can be mapped to more entities from the text, we select the mapping which maximizes global fitness.
The global fitness value is then used to determine the relation between text–hypothesis pairs. The "No entailment" cases are represented by pairs for which the global fitness value is below a threshold, the value of which is extracted from the training data, and the "Entailment" cases are represented by pairs for which global fitness is above the same threshold; for separating contradiction and unknown cases, we considered another threshold, also extracted from the training data.

3.1. Entailment Cases

3.1.1. Basic Positive Rules

In order to determine the global fitness for a given pair, we need a mapping of the nodes of the hypothesis tree to the nodes of the text tree. 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 direct mapping is not possible, we use external knowledge bases to transform the hypothesis node into a version that is more similar to some node in the text. For *verbs* we use DIRT (Lin and Pantel, 2001) and transform the hypothesis tree into an equivalent one, where the verb node is replaced with an equivalent form. This is the case of pair 538 where in the text we have "*Long John Baldry, English-born blues legend, passed away…*" and in the hypothesis we have "*A musician has died…*". After using this resource, the hypothesis has changed in "*Obama passed away*" 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.
In the case of *named entities*, we either use an acronym database[4] or obtain information related to it from background knowledge (Iftene and Balahur, 2008). An example for acronyms is found in pair 72 where in the text we have *United States* and in hypothesis we have *US*. Examples in which we use our module which adds new elements from English Wikipedia to the background

[3] http://www.alias-i.com/lingpipe/
[4] http://www.acronym-guide.com

knowledge are pairs 364 (relation between *Basel in Switzerland* and *European city*), 424 (between *Canadian Prime Minister* and *Prime Minister of Canada*), etc.
For *nouns* and *adjectives* we use WordNet (Fellbaum, 1998) and some of the relations from eXtended WordNet[5] to look up synonyms, which we then attempt to map to nodes from the text tree. Examples of synonymy relation from WordNet are pairs 17 (between *blocked* and *ban*), 26 (between *holds* and *detained*), 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 case of numerical data, some special situations need to be taken into account. There are cases in which, even if the numbers from the text and the hypothesis are not the same, certain quantifiers may change their meaning enough for a positive match. For solving these cases, we create intervals for both expressions and since the interval from the text is contained in the interval from the hypothesis, we award a local fitness value of 1. The quantifiers are taken from a list which contains expressions like "more than", "less than", or words such as "over", "under", etc.

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 a 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 is negated or not. For determining negation, we check inside the verb's subtree tree to see whether words such as "not", "never", "may", "might", "cannot", etc appear. For each of these words we successively negate the initial truth value of the verb, which by default is "false".
A specific rule was also built for the particle "to" preceding 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.

[5] http://xwn.hlt.utdallas.edu/

3.2.3. Contradiction Cases

For determining contradiction, we consider several situations, most common of which is the negation of the verb with words like "never", "not", "no", "cannot", "unsuccessfully", "false" etc. This case is encountered at pair 522 where in the text we have "*Movie studio company, New Line Cinema has announced that movie director Peter Jackson will never be allowed to work on another New Line film.*" and in the hypothesis "*New Line wants to work with Peter Jackson.*".
Another contradiction case is that of long infinitive verbs preceded by words such as "refuse", "deny", "ignore", "plan", "intend", "proposal", "able", etc.
Contradiction is also determined by identifying the antonymy relation between words from the text and the hypothesis. For finding antonymy we use the [*opposite-of*] relation from VerbOcean (Chklovski and Pantel, 2004) and antonymy relation from WordNet. In order to broaden the domain of the antonymy relation, we consider a combination of synonyms and antonyms from WordNet or opposites from VerbOcean. For words from the hypothesis which cannot be mapped to words from the text using either synonymy or antonymy, we consider the set of antonyms for their synonyms and then check if any word from this new set cans me mapped to the text.
In some situations, the similarity relation from DIRT is an antonymy relation, and for this reason we do an extra verification of DIRT relations to see if we have antonymy in either WordNet or VerbOcean. For all identified *contradiction cases*, since we consider the penalties with the highest values, the final answer for the considered pairs will be "*Contradiction*".

3.2.4. Unknown Cases

If the text or hypothesis contains words such as "may", "can", "should", "could", "must", "might", "infrequent", "rather", "probably", etc., the penalties are not decisive in establishing the final answer, which is obtained only after computing global fitness. At pair 46 we have in the text "…*could eventually be taken over* …" and in hypothesis we have "… *is* …".
With regards to the particle "to" we will consider those cases which are not determined to be contradictions.
In the case of *named entities*, however, the solution we have chosen is different. If even after using the acronym database we cannot map the entity from the hypothesis to an entity in the text, we decide that a pertinent conclusion cannot be drawn, and the result for the pair is "*Unknown*". This case is found at pair 34 given bellow, where we have the named entity *UK* in the hypothesis without a corresponding value in the text.

> T: *Speaking after he discovered that he would not face criminal charges, Mr Green disclosed that the officers who arrested him last November warned him that he could be given the longest possible sentence. "They said, 'Do you realise that this offence could lead to life imprisonment?'," Mr Green told BBC Newsnight. I just thought this was absurd. "I assume it's because it's a common law offence therefore because there is no law on the statute book which I was alleged to have broken, then presumably there is no set sentence for it."*
>
> H: *Mr. Green is the shadow immigration minister of the UK.*

If any of the numbers in the text or the hypothesis has an attached unit of measure, it is always kept, as it is possible to find the same numbers in the text and the hypothesis, but to have those numbers referring to different entities:

T: *At least 14 people have been killed in a suicide bomb attack in southern Sri Lanka, police say. The telecoms minister was among about 35 people injured in the blast at the town of Akuressa, 160km (100 miles) south of the capital, Colombo. ...*

H: *35 government officials were injured by a suicide bomber in Akuressa.*

An exception to the named entity rule presented above is the case when the entity is a *first name*, in which case we only insert a penalty in the global fitness:

T: *... The man accused of killing Ms. Zapata, Allen R. Andrade, 32, told the police that he had attacked her upon discovering that she was biologically a man, after the two met on the Internet and had a sexual encounter. In a chilling arrest affidavit, Mr. Andrade said he thought he had "killed it," after striking Ms. Zapata in the head until she stopped breathing. ...*

H: *Angie Zapata has been killed with a fire extinguisher.*

4. Results in RTE5

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

Answer Type	**In Gold**	**Correct offered by our system**	**Total offered by our system**	**Precision**	**Recall**	**F-measure**
Entailment	300	260	379	68.60%	**86.67%**	**76.58%**
Contradiction	90	22	44	50.00%	24.44%	32.84%
Unknown	210	128	177	**72.32%**	60.95%	66.15%
Total	**600**	**410**	**600**	**68.33%**		

Table 1: Results in RTE5 on 3-way task on test data

As can be seen in the analysis of the results, our system does worst on the *Contradiction* cases and best on the *Entailment* cases. This is similar to the results we obtained for the RTE5 training set, given in Table 2 below.

Answer Type	**In Gold**	**Correct offered by our system**	**Total offered by our system**	**Precision**	**Recall**	**F-measure**
Entailment	300	266	382	69.63%	**88.67%**	**77.22%**
Contradiction	90	16	36	44.44%	17.78%	25.40%
Unknown	210	137	182	**75.27%**	65.24%	69.90%
Total	**600**	**419**	**600**	**69.83%**		

Table 2: Results in RTE5 on 3-way task on training data

For the 2-way task, the distribution is presented in table below:

Answer Type	**In Gold**	**Correct offered by our system**	**Total offered by our system**	**Precision**	**Recall**	**F-measure**
Yes	300	260	379	68.60%	86.67%	76.58%
No	300	181	221	**81.90%**	60.33%	69.48%

Total	**600**	**441**	**600**	**73.50%**		

Table 3: Results in RTE5 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 (81.9%), where from 221 answers offered by our system 181 are correct. The meaning of the difference between global precision from 2-way task and 3-way task is that in 31 out of 221 cases we don't distinguish correctly between *Contradiction* and *Unknown* cases.
According to the source of test data, we can see in Table 4, how in RTE5 we got comparable results with results from RTE4, with an improvement for data from QA task. In comparison with results from RTE3 we can see that we have significant improvements on IR and IE tasks, but on for QA task, where we got the best results in RTE3, and worst result in RTE4.

Provenience of testing data	**RTE3**	**RTE4**	**RTE5**
IR	69.00 %	82.00 %	84.0 %
QA	87.00 %	63.00 %	70.5 %
SUM	63.50 %	78.00 %	*Na*
IE	57.00 %	64.33 %	66.0 %
Total	**69.13 %**	**72.10 %**	**73.5 %**

Table 4: Comparison between results between RTE3, RTE4 and RTE5

5. Ablation Tests

Following the RTE-3 competition in order to determine each component's relevance, the system was run in turn with each component removed (Iftene, 2009). The same technique was used after that in RTE-4 and in RTE-5. Table 4 presents these results in parallel for RTE-3, RTE-4 and RTE-5, where the meanings for P, C and WR are: P = *Precision*, C = *Contribution* and WR = *Weighted Relevance*.

System Description	**RTE-3 (69.13 %)**			**RTE-4 (72.1 %)**			**RTE-5 (73.5 %)**		
	P (%)	C (%)	WR (%)	P (%)	C (%)	WR (%)	P (%)	C (%)	WR (%)
Without DIRT	68.76	0.37	0.54	71.40	**0.7**	**0.97**	73.33	0.17	0.23
Without WordNet	68.00	1.13	1.63	69.10	**3.0**	**4.16**	72.5	1.00	1.36
Without Acronyms	68.38	**0.75**	**1.08**	71.80	0.3	0.42	73.33	0.17	0.23
Without BK	67.75	1.38	2.00	70.40	**1.7**	**2.36**	72.33	1.17	1.59
Without the NE rule	57.58	**11.55**	**16.71**	66.90	5.2	7.21	67.33	**6.17**	**8.39**
Without the Negation rule	67.63	1.50	2.17	68.70	**3.4**	**4.72**	73.5	0.00	0.00
Without the Contradiction rule	-	-	-	68.10	**4.0**	**5.55**	71.5	**2.00**	**2.72**
Without additional	-	-	-	-	-	-	69.33	**4.17**	**5.67**

System Description	RTE-3 (69.13 %)			RTE-4 (72.1 %)			RTE-5 (73.5 %)		
	P (%)	C (%)	WR (%)	P (%)	C (%)	WR (%)	P (%)	C (%)	WR (%)
processing steps									
Total		16.68	24.13		18.3	25.39		14.85	20.20

Table 5: Components' relevance for 2-way task

The meanings of the columns for RTE-3 competition are the following (similar for RTE-4 and RTE-5 columns):

- $Precision_{Without\_Component}$ value was obtained running the RTE-3 system without a specific component (for example, $Precision_{Without\_DIRT}$ is 68.76 % and it represents the precision of the RTE-3 system without the DIRT component);
- $Contribution_{Component} = Full\_system\_precision - Precision_{Without\_Component}$ (for example, $Contribution_{DIRT}$ is 69.13 % - 68.76 % = 0.37 % for the DIRT component of the RTE-3 system, where 69.13 % is the precision for the full RTE-3 system and 68.76 % is the precision for RTE-3 system without DIRT component);
- $WeightedRelevance_{Component} = \frac{100 \times Contribution_{Component}}{Full\_system\_precision}$ (for example, for the DIRT component in RTE-3, $WeightedRelevance_{DIRT} = \frac{100 \times Contribution_{DIRT}}{Full\_system\_precision} = \frac{100 \times 0.37}{69.13} = 0.54\%$).

The results in Table 5 show that the system's rules related to negation, named entities and contradictions are the most important. In RTE-5 we also perform ablation test for the module related to the additional processing steps that include preparation of input data, identification of named entities, with our patterns or using GATE, and running of TreeTagger.

The Table 6 presents a comparison between ablation tests performed on RTE-5 data for 2-way and 3-way:

System Description	2-way (73.5 %)		3-way (68.33 %)	
	P (%)	**C (%)**	**P (%)**	**C (%)**
Without DIRT	73.33	0.17	68.00	0.33
Without WordNet	72.50	1.00	67.00	1.33
Without Acronyms	73.33	0.17	68.17	0.17
Without BK	72.33	1.17	66.83	1.50
Without NE rule	67.33	**6.17**	63.33	**5.00**
Without the Negation rule	73.50	0.00	66.83	1.50
Without the Contradiction rule	71.50	**2.00**	69.67	- 1.34
Without additional processing steps	69.33	**4.17**	64.33	**4.00**
Total		**14.85**		**12.49**

Table 6: Components' relevance in RTE5

We can see in the table above the importance of the resources used for 2-way and 3-way tasks. It is interesting to see that one of the most valuable rules from last year's system, the rule that identifies contradictions, has a negative contribution to the overall result of the this year's system for the three way task.

6. Pilot Task

RTE-5 introduced a pilot task, concerning the extraction of text from a series of newspaper articles that yielded positive entailment for a given set of hypotheses. The difficulty of the task is twofold: first, the texts are not modified in any way as compared to the original source, so they may contain spelling errors, sentences with grammar errors, abbreviations and contractions, etc. The second problem is that there are a large numbers of candidate pairs, as for every one of the nine topics there are about ten hypotheses, and for every hypothesis in a topic the number of candidate pairs is equal to the number of sentences. This leads to a very large search space, and the problem to reduce it becomes very important.

In order to reduce the search space, we have made use of a technique used for our question answering systems, described in (Iftene et al., 2009). First, using Lucene, we have indexed the articles from each topic at the sentence level, thus obtaining nine indexes. Then we have built queries for all the hypotheses by removing all punctuation and stop words, which we then used to extract the relevant text snippets. Based on experiments on the training data, we have determined that the snippets with the highest chance of yielding positive entailment are clustered around the top scoring snippets, and the first item that is not in the cluster has a Lucene score at least three times lower than that of the last item in the cluster. We have also empirically determined that the smallest feasible number of candidates is ten, and that a candidate number of above twenty is too large. In practice, the number of candidates selected is almost always above fifteen.

In order to determine the entailment value of the candidate pairs (approximately 1700 in all), we have applied a lightweight version of our entailment system. The results are given in Table 7 below:

Result	**Precision**	**Recall**	**F-measure**
Micro-average	51.12%	22.88%	31.61%
Macro-average Topic	53.03%	24.08%	33.12%
Macro-average Hypothesis	46.55%	26.42%	33.71%

Table 7: Results for RTE-5 pilot task

7. Conclusions

The paper presents the architecture of the system used in RTE-5. This system is an improved version of the system used in RTE-4 (Iftene, 2008), and has new important components. First, we transform the input data in a new format. The second one of the new components is responsible with identification of named entities of type Job and Language. The last one tries to correct named entities from the hypothesis without correspondence in the text, using the Google Search API and "*Did you mean*" option.

With the new changes presented, in comparison with system used in RTE-4, the precision increased by 1.4 % for 2-way task, and the precision decreased with 0.17 % for 3-way task.
The main problems are related to cases in which text contains almost all words from the hypothesis, in the same order, but the constituents have different semantic roles.

Acknowledgments

The authors thank 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.

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. *UAIC Participation at RTE4*. In Text Analysis Conference (TAC 2008) Workshop - RTE-4 Track. National Institute of Standards and Technology (NIST). November 17-19, 2008. Gaithersburg, Maryland, USA.

Iftene, A. 2009. *Textual Entailment*. PhD Thesis. "Al. I. Cuza" University. March 13, 2009. Iasi, Romania. (http://thor.info.uaic.ro/~adiftene/thesisAI.pdf)

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.

Iftene, A., Trandabăţ, D., Pistol, I., Moruz, A., Husarciuc, M., Cristea, D. 2009. *UAIC Participation at QA@CLEF2008*, Evaluating Systems for Multilingual and Multimodal Information Access, Lecture Notes in Computer Science, vol. 5706/2009, pp. 385-392, ISBN 978-3-540-74998-1, ISSN 0302-9743 (Print) 1611-3349.

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.