Answer Validation on English and Romanian Languages

Adrian Iftene[1], Alexandra Balahur-Dobrescu[1, 2]

[1] UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
[2] University of Alicante, Department of Software and Computing Systems, Alicante, Spain
{adiftene, abalahur}@info.uaic.ro

Abstract. The present article presents the steps involved in the transformation of the TE system that was used in the RTE3 competition in 2007 for the AVE 2008 exercise. We describe the rules followed in building the patterns for question transformation, the generation of the corresponding hypotheses and finally for answer ranking. We conclude by presenting an overview of the performance obtained by this approach and a critical analysis of the errors obtained.

1 Introduction

AVE[1] (Answer Validation Exercise) is a task introduced in the QA@CLEF competition, with the aim of promoting the development and evaluation of subsystems validating the correctness of the answers given by QA systems. Participant systems receive a set of triplets (Question, Answer, and Supporting Text) and they must return a judgment of SELECTED, VALIDATED or REJECTED for each triplet.

This year, for our second participation in the AVE competition, we improved the system used last year and, additionally introduced a question analysis part, which is specific to a question answering system. In this year's AVE competition we also participated with a system working in Romanian, using a Textual Entailment (TE) system working on Romanian. The latter is similar to the TE system working in English with which we participated in the RTE 3 competition in 2007 (Iftene, Balahur-Dobrescu, 2007b). Due to this reason, the present paper describes solely the AVE system working in English. The following sections present the new functionalities that have been added to our English TE system.

2 Textual Entailment System

The main architecture of our Textual Entailment system remains the same (Iftene, Balahur-Dobrescu, 2007a). The goal of the system is to transform the hypothesis making use of extensive semantic knowledge from resources like DIRT, WordNet, Wikipedia, and database of acronyms. Additionally, we built a system to acquire the extra Background Knowledge needed and applied complex grammar rules for rephrasing in English. Tools used are LingPipe[2] and MINIPAR[3] (Lin, 1998).

Based on a tree edit distance algorithm (Kouylekov and Magnini, 2005), the main goal of our algorithm is to map every entity in the dependency tree associated with the hypothesis to an entity in the dependency tree associated with the text.

For every mapping, we compute a local fitness value, which indicates the appropriateness between entities. Based on this local fitness, further on an extended local fitness is computed and, eventually, using all partial values, the global fitness is summed up. Two rules are also added for the global fitness calculation, namely:

- The *Semantic Variability Rule* – which is a rule regarding the negation of verbs, words that are "stressing certainty (preserving it)" regarding the sense of the sentence and, on the other hand, words that are "certainty diminishing" the sense of the sentence, negating it;
- The *Rule for Named Entities* - The rule is applied for named entities from the hypothesis which have no correspondence in the text. If the word is marked as named entity by LingPipe, we try to use the acronyms' database or obtain information related to it from the background knowledge. In last year's

[1] http://nlp.uned.es/QA/ave/
[2] http://www.alias-i.com/lingpipe/
[3] http://www.cs.ualberta.ca/~lindek/minipar.htm

version of the TE system, in the event that even after these operations we cannot map the word from the hypothesis to one word from the text, we set the value for the global fitness to 0.

The main change from this year is regarding the Rule *for Named Entities*. There are cases where it is possible for all pairs to have hypotheses with Named Entity problems. With the old rule, the global fitness for all these pairs is set to 0, and it is thus impossible to select the best value (corresponding to the SELECTED answer).

In the new rule, we compute the global fitness value for current pair, but we also mark the current pair as having a "NE Problem". Further on, we will see how this marking is used in ordering the answers and in the final evaluation for the AVE task.

Changing this rule helps our program in cases such as that of the question with id = "0054":

Table 1: Question with id = "0054"

In what date did Mathieu Orfila write his "Traité des poisons"?

In which all justification snippets for answers contain the name "*Mathieu Orfila*", but don't contain the exact article name "*Traité des poisons*". From all possible answers, we select as correct the answer "1813", with the justification snippet:

Table 2: Justification snippet for question with id = "0054"

Mathieu Orfila is considered to be the modern father of toxicology, having given the subject its first formal treatment in 1813 in Mathieu Orfila Trait des poisons, also called Toxicologie generate.

3 Using the TE System in the AVE track

The system architecture for this year is presented below:

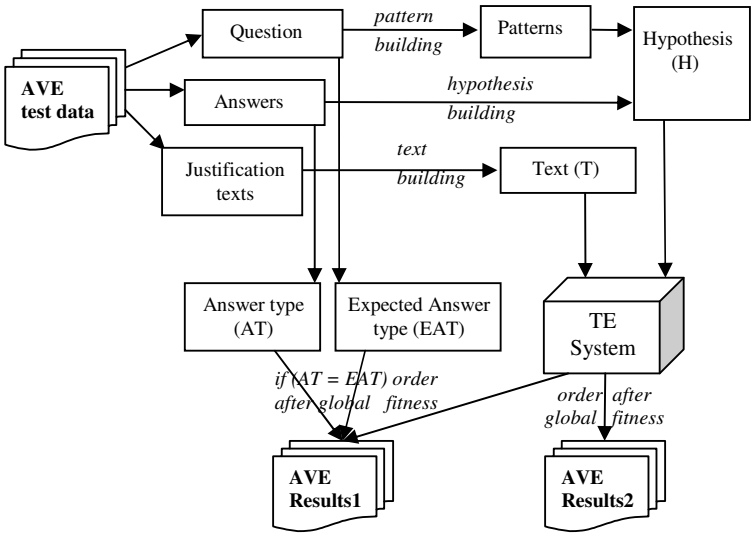

Figure 1: AVE System. The structure is similar for English and Romanian

The steps executed by our system are the following:

- From the system built for AVE 2007, we keep the following steps:
  - We build a pattern with variables for every question according to the question type;
  - Using a pattern and all possible answers, we build a set of hypotheses for each of the questions: H_1, H_2, H_3 etc.;
  - We assign the justification snippet the role of text T and we run the TE system for all obtained pairs: (T_1, H_1), (T_2, H_2), (T_3, H_3), etc.
- Additionally, we perform the next steps:
  - Identify for the answers the Answer Type (AT);
  - Identify for the questions the Expected Answer Type (EAT).

Lastly, we submit two results for our system:

1. In the first one we consider the correct answer for the current question the candidate from the hypothesis for which we obtain the greatest global fitness;
2. In the second one, we consider the correct answer for the current question the candidate with AT equal with EAT and for which we obtain the greatest global fitness.

3.1 Pattern Building

In order to use the TE system for ranking the possible answers in the AVE task, all these questions are first transformed according to the algorithm presented in (Bar-Haim et al., 2006).

For question 13 we have:

Question: *What is the occupation of Richard Clayderman*?

Our program generates the following pattern:

Pattern: *The occupation of Richard Clayderman is JOB.*

where *JOB* is the variable in this case. We generate more specific patterns this year according to the following answer types: City, Count, Country, Date, Job, Measure, Location, Person, Organization, Year and Other. Next table presents the identified types of patterns:

Table 3: Examples of Patterns

Answer type	Cases Number	Question example	Pattern
City	3	What is the capital of Latvia?	The capital of Latvia is **CITY**.
Count	14	How many "real tennis" courts are there?	**COUNT** "real tennis" courts are there.
Date	10	When was Irish politician Willie O'Dea born?	Irish politician Willie O'Dea was born **at DATE**.
Job	4	What is the occupation of Jerry Hickman?	The occupation of Jerry Hickman is **JOB**.
Measure	14	What distance is run in a "Marathon"?	**MEASURE** is run in a "Marathon".
Location	14	Where does Muriel Herkes live?	Muriel Herkes lives **in LOCATION**.
Person	17	Which composer wrote "pacific 231"?	**PERSON** wrote "pacific 231".

Answer type	Cases Number	Question example	Pattern
Organization	14	What is the political party of Tony Blair?	**ORGANIZATION** is the political party of Tony Blair.
Year	6	In what year did Emerson Lake&Palmer form?	Emerson Lake&Palmer was form **in YEAR**.
Other	21	In Japanese, what is "bungo"?	"bungo" is **OTHER**.

Following the building of the pattern, we proceed to constructing the corresponding hypotheses. A special case is for DEFINITION questions, when we didn't build any pattern (in this case only the answer will be the hypothesis).

3.2 Hypothesis building

Using the pattern building mechanism above and the answers provided within the AVE data, we built the corresponding hypotheses. For example, for question 27, we build, according to the answers from the English test data ("**a_str**" tags), the following hypotheses:

H_{13_1}: *The occupation of Richard Clayderman is Number.*

H_{13_2}: *The occupation of Richard Clayderman is teacher Qualifications.*

H_{13_3}: *The occupation of Richard Clayderman is ways.*

H_{13_8}: *The occupation of Richard Clayderman is pianist.*

H_{13_11}: *The occupation of Richard Clayderman is artist.*

H_{13_12}: *The occupation of Richard Clayderman is Composer.*

H_{13_13}: *The occupation of Richard Clayderman is teachers.*

For each of these hypotheses, we consider as having the role of text *T* the corresponding justification text (content of the "**t_str**" tag).

3.3 Global Fitness Calculation

We consider the pairs built above as input for our Textual Entailment system. After running the TE system, the global fitness values and the values with marked "NE Problems" for these pairs are the following:

Table 4: TE System output

Pair	Global Fitness	NE Problem
13_1	1.5	Clayderman
13_2	2.35	Clayderman
13_3	2.31	Clayderman
13_8	1.92	
13_11	1.82	
13_12	1.86	
13_13	1.89	Clayderman

3.4 Answers Type and Expected Answer Type Identification

The aim in performing this step is to eliminate the cases in which there are differences between these values. For example, in the case of question 13, since the expected answer type is JOB, it is normal to try to identify the correct answer in the sub-set of answers of type JOB.

The patterns used in the identification of the expected answer type (EAT) are similar to the patterns used in 3.1. For the identification of the answer type (AT), we use GATE[4] for the following types: Job, City, Country, Location, Person, Organization and we build specific patterns in order to identify the following types: Date, Measure, and Count. When an answer cannot be classified with GATE or with our patterns, it is considered with type Other. For question number 13, we have:

Table 5: EAT and AT comparison

Pair	EAT	Answer	AT	Match score
13_1	JOB	Number	OTHER	0.25
13_2	JOB	teacher Qualifications	OTHER	0.25
13_3	JOB	Ways	OTHER	0.25
13_8	JOB	Pianist	JOB	1
13_11	JOB	Artist	JOB	1
13_12	JOB	Composer	JOB	1
13_13	JOB	teachers	JOB	1

On last column is the matching score between EAT and AT. In order to compute this value, we use a set of rules. The most important rules are:

Table 6: Rules for matching score calculation

Rule	Match score
AT = EAT	1
(EAT = “DEFINITION”) and (AT = “OTHER”)	1
EAT and AT are in the same class of entities: {CITY, COUNTRY, REGION, LOCATION} or {YEAR, DATE} or {COUNT, MEASURE, YEAR}	0.5
(AT = “OTHER”) or (EAT = “OTHER”)	0.25
OTHERWISE	0

3.4 Answers classification

We submit two runs on each of the languages (English and Romanian) according to the use or not of some system components. The systems are similar and only the external resources used by the TE system or by GATE are language-specific.

First run: is based on TE System output. The answers for which we have NE problems are considered as REJECTED (for question 13, using table 4, we can deduce that answers 1, 2, 3 and 13 are REJECTED). Answers

[4] http://www.gate.ac.uk/

without NE problems are considered as VALIDATED (answers 8, 11, 12) and the answer with the highest global fitness is considered as SELECTED (answer 8). If all answers contain NE problems, then all answers are considered REJECTED, except the answer with highest global fitness, which will be considered SELECTED.

Second run: in addition to the first run, we add the comparison between EAT and AT. In the cases where we have NE Problems, the answers are considered as REJECTED as well, and we also take into consideration if the matching score between EAT and AT is 0 (incompatible types). Of the remaining answers, if the matching score is not 0, then all answers are VALIDATED. For the identification of the SELECTED answer, we select the answers with the highest matching score (8, 11, 12) and the highest global fitness. In this case, the results are the same.

3.5 Results

Our AVE systems have the following results:

Table 7: AVE Results in 2008

	English		Romanian	
	Run1	**Run2**	**Run1**	**Run2**
F measure	0.17	0.19	0.22	0.23
Precision over YES pairs	0.09	0.11	0.12	0.13
Recall over YES pairs	0.76	0.85	0.92	0.92
QA accuracy	0.19	0.24	0.17	0.24
Estimated QA performance	0.19	0.24	0.17	0.25

Table 8: Distribution of our results on answer classes for Run2 in English

Answers Class in Gold file	**Unknown**	**Validated**	**Rejected**	**Total**
Correct	0	67	398	**465**
Incorrect	36	49	542	**627**

4 Conclusions

Last year, we showed how the TE system used in the RTE3 competition can successfully be used as part of the AVE system, resulting in improved ranking between the possible answers, especially in the case of questions with answers of type Person, Location, Date and Organization. This year, changing some of the rules employed in the Textual Entailment system and adding the question and answer type classification and matching component, we showed how we improved, on the one hand, the correct classification of the answers, and on the other hand, the validation of more answers.

One of the main problems encountered was the class of UNKNOWN answers types, which our system does not identify. In order to detect these cases, a three way classification of the answers, such as that proposed in the RTE 3 pilot task is intended to be used in the future.

The rule regarding the presence of NEs remains of great importance, identifying the correct cases. However, in the cases where the NER or NEC is not correctly performed, the system fails. Moreover, this rule is not enough to identify the entire class of REJECTED answers, as shown in table 8. In order to better identify these situations, additional rules must still be explored in order to further improve the system.

References

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini B., Szpektor, I. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment. Venice. Italy.

Kouylekov, M., Magnini, B. 2005. Recognizing Textual Entailment with Tree Edit Distance Algorithms. In Proceedings of the First Challenge Workshop Recognising Textual Entailment, Pages 17-20, 25–28 April, 2005, Southampton, U.K.

Iftene, A., Balahur-Dobrescu, A. 2007a. 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. 28-29 June, Prague, Czech Republic.

Iftene, A., Balahur-Dobrescu, A. 2007b. Improving a QA System for Romanian Using Textual Entailment. In Proceedings of RANLP workshop "A Common Natural Language Processing Paradigm For Balkan Languages". Pages 7-14, September 26, Borovets, Bulgaria.

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