Adaptation of a Machine-learning Textual Entailment System to a Multilingual Answer Validation Exercise

Zornitsa Kozareva, Sonia Vázquez and Andrés Montoyo

NLP Research Group (GPLSI)
University of Alicante, Spain
{zkozareva,svazquez,montoyo}@dlsi.ua.es

Abstract. The recognition of textual entailment is a new generic task in which a system automatically establishes whether a given text entails the meaning of another. This detection is useful for many NLP applications. In the case of Question Answering, textual entailment is represented as an answer validation exercise (AVE) where the system has to verify the correctness of the returned snippet of the Question Answering system. In this paper we present an AVE system based on the combination of word overlap and Latent Semantic Indexing modules. The main focus and contribution of our work consist in the adaptation of the already developed and evaluated textual entailment system MLEnt and its ability to function with different languages. We have evaluated our approach with the English, Spanish, French, Dutch, German, Italian and Portuguese languages.

Keywords: Answer Validation (AV), Question Answering (QA), Recognising Textual Entailment (RTE), Multilingual

1 Introduction

Natural Language Processing (NLP) has different tasks such as Information Retrieval (IR), Word Sense Disambiguation, Automatic Summarisation, Question Answering (QA) among others. Each task is associated with different problems: ambiguity, variability, different language characteristics, specific domain, specific language and for their resolution different techniques are used. Thus, in order to solve a given problem current researchers use different approaches and systems. However, all of these systems have to be evaluated. In most cases, to measure the effectiveness of an IR or QA system, it is necessary to use humans. For a human this task is labour-intensive and time-consuming. Therefore, nowadays researchers study different ways through which this evaluation process can become automatic.

In consequence of the introduction of the textual entailment recognition (RTE)[1] [2], [3] task whose aim is to determine whether the meaning of one text is entailed by the meaning of another text, the answer validation exercise (AVE) [7] emerged. An AVE system is centered on the validation of the QA question which is transformed into affirmative mode and the snippet returned by the QA system. The AVE pair is considered correct only if the QA question and snippet are correct.

Since RTE and AVE tasks are related, we decided to evaluate our already developed RTE system MLEnt [5] in the AVE challenge. The attributes of MLEnt do not use any language dependent tools or resources, which means that the system is able to operate on different languages. However, there is no multilingual textual entailment corpora and we could not prove our claims. In the AVE challenge the multilingual data is present, therefore we decided to evaluate our approach and confirm our hypothesis of MLEnt's multilinguality.

In the AVE challenge, our research focuses on two aspects. The first one is the portability of our text entailment system to the AVE challenge together with its issue of multilinguality. This

[1] http://www.pascal-network.org/Challenges/RTE/
http://www.pascal-network.org/Challenges/RTE2/

aspect is proven with the exhaustive evaluation on the following seven languages: English, Spanish, French, Dutch, German, Italian and Portuguese. The second aspect of our study is the influence of voting. For the RTE challenge [4] we have demonstrated that depending on the data sets the voting strategy has higher or almost no impact. This made us observe the voting behaviour for the AVE challenge.

The paper is organized as follows: Section 2 describes the different modules of our AVE system, in Section 3 are shown the carried out experiments, followed by discussion of the obtained results and finally we conclude in Section 4.

2 System Description

In this section, we describe our answer validation system. The system consists of a module of the already developed Machine-learning based Textual Entailment system called MLEnt. This module is combined through simple voting technique with a newly developed module for acquiring semantic similarity information through Latent Semantic Indexing. The developed system does not use any language dependent tools or resources, therefore it was possible to evaluate its performance for different language answer validation pairs. The main advantages of our system are its low computational cost, fast performance and the ability to work on different languages.

2.1 Word overlap module

The MLEnt system consists of word overlap and semantic similarity modules. For the AVE competition we utilized only the word overlap module as the similarity one uses information from WordNet and it was impossible to adapt it to the different language pairs. The following attributes were included in the AVE word overlap feature set:

n-grams: look for common position-independent unigram matches between the text and the hypothesis in the AV pair. According to this attribute the AV pair is correct when the sentences of the text and the hypothesis share the same words. Respectively, the n-gram attribute determines the AV pair as incorrect when there are no common words at all.

This attribute does not consider semantic similarity information and it is not be able to determine that if "vehicle" and "car" are present in the sentences they are related and this can increase the similarity of the n-gram approach. Another obstacle for this attribute is its insensitiveness to the word order and the sentence level structure. As it is looking only for arbitrary n-gram coincidences, it is unable to determine that although "Mary calls the police" and "the police calls Mary" have the same words, they do not infer the same meaning. For this reason, we included the longest common subsequence (LCS) and the skip-gram attributes.

LCS: looks for common non-consecutive word sequences of any length between the sentences of the text and the hypothesis. Longer LCS corresponds to more similar sentences for the answer validation text and hypothesis. LCS estimates the similarity between text T with length m and hypothesis H with length n, as $\frac{LCS(T,H)}{n}$. LCS does not require consecutive matches but in-sequence matches that reflect the sentence level word order. When the measure determines two or more equal by length LCS, only one of them is considered. For this measure there is no need to define the length of the word sequences, because LCS automatically includes the longest in-sequence n-gram. LCS reflects the proportion of ordered words found in the text and the hypothesis, therefore in comparison to the n-grams measure it will indicate that "Mary calls the police" and "the police calls Mary" are not so similar because they share only two ordered n-grams. This influences the AV classification in a more sensitive way denoting that the text and the hypothesis have grater distance than the one determined by the n-grams.

skip-gram: represent any pair of words in sentence order that allow arbitrary gaps. Once all pairs for the text and the hypothesis are generated, the overlapping skip-grams are counted using the measure $skip\_gram = \frac{skip\_gram(T,H)}{C(n, number\_of\_skip\_gram)}$, where $skip\_gram(T,H)$ refers to the number of common skip-grams found in the text T and the hypothesis H, $C(n, number\_of\_skip\_gram)$ is a combinatorial function, where n is the number of words in the hypothesis and $number\_of\_skip\_grams$

corresponds to the number of common n-grams between (T, H) [2]. According to the skip-grams, the AV pairs is correct when text and hypothesis have more common skip-grams.

For the following sentence pairs:

S_1: Mary calls the police.

S_2: Mary called the police.

S_3: The police called Mary.

the skip-grams identify that the similarity between the sentences S_1 and S_2 is stronger than the similarity between the sentences S_1 and S_3 or S_2 and S_3. However, the n-gram and the LCS are not so sensitive and cannot determine the similarity correctly.

number matching: identifies the numbers in the AV text and hypothesis, and then verifies them. For sentences where there are no numbers at all, the number matching attribute assigns the value of NO to the AV pair. According to this attribute, the AV is correct when the numbers in the text and the hypothesis coincide.

The performance of the described attributes is evaluated only for English and Spanish, because the AVE organizers provided training data for these languages. For the training phase we used the SVM and kNN machine learning classifiers, and also observed the behaviour of the information gain (IG) measure for the different language pairs and for different sizes of training data. IG is a measure that indicates from a given set of features which are the most important ones. According to IG, the two most crucial attributes for the AVE correct classification are the LCS and the skip-gram. For the word overlap feature set, the system generated two outputs, one obtained by the LCS and another obtained by the skip-gram.

We had to adjust the LCS and skip-grams attributes to the rest of the languages for which we had no training data. As the generated attributes are influences by the length of the overlapping words normalized by the total amount of words present in the AV hypothesis it was possible to adapt them. We measured the standard deviation of the LCS and the skip-gram attributes for Spanish and English, and we observed the obtained standard deviation values and their corresponding YES/NO values. Thus for the rest of the languages, the YES/NO values were assigned according to the standard deviation values.

2.2 Latent Semantic Indexing module

Latent Semantic Indexing (LSI) [6] is a computational model that establishes the relations among the words in a large corpus using a vectorial-semantic space where all terms are represented by a term-document matrix or so called conceptual matrix. In order to obtain the similarity relations, the terms have to be distributed in documents, paragraphs or sentences. Then according to this distribution, the co-occurrence among the different terms is determined.

Once the term-document matrix is obtained, LSI uses the recursive algorithm Singular Value Decomposition (SVD) which decomposes the term-document matrix into three other matrices. These matrices contain the singular vectors and the singular values. SVD transforms the original data into linearly independent factors. Many of these factors are very small and they can be ignored in the approximation model. The final result of the decomposition process is a reduced matrix of the initial term-document matrix that is used to establish the word similarities.

In order to apply LSI to the AVE task, we need corpus which serves as a basis to construct the conceptual matrix. We used the answer validation T-H data provided by the AVE organizers. The conceptual matrix is constructed with the T-phrases of the AVE corpus. This is due to the results which we obtained in a study with the RTE2 data [9]. According to this study when the T phrases are used as corpus, the RTE task performs better. For each one of the languages – English, Spanish, Italian, German, Dutch, Portuguese and French, we constructed different conceptual matrices using the sentences of the text from the AVE corpora.

From the generated conceptual matrix one can establish the similarities of the terms, the phrases or the documents. In our experiment, we are interested in establishing the similarity of

[2] (e.g. *number_of_skip_grams* is 1 if there is a common unigram between T and H, 2 if there is a common bigram etc.)

the T-H pairs. LSI extracts the similarity relations between the T-H phrases and the results is a list of T-H phrases ordered by their similarity score. Under the concept of LSI, an AVE pair is correct when the similarity value is close to 1, and less similar when the similarity value is close to 0.

In order to determine which values should be considered as more or less similar, we used a threshold. This threshold was determined after we executed several experiments with the Spanish and English AVE training data. For the both languages the best results were obtained with the 0.8 threshold. Which means that a T-H pair which equals or is above 0.8 is assigned with the value 'YES' and the rest of the pairs are assigned with 'NO' value. The 0.8 value was also used as a threshold for the rest of the languages.

The next examples illustrate how the LSI module works:

Example1: This is an instance of the AVE test collection that returns an entailment value of 'NO', in this case the LSI Module returns a value of 0.402886:

< *pair* id="4525" value="NO" task="QA">

< *q* lang="EN" src="clef2006" type="OBJECT">**What is Atlantis**< */q* >

< *t* doc="096222">**TO ATLANTIS' CREW. From Associated Press NASA briefly lost contact with the space shuttle Atlantis and its six astronauts Sunday because of crossed radio signals. The problem occurred as Atlantis switched from one Tracking and Data Relay Satellite to another, a routine procedure during Atlantis nor its crew was in any danger, and no science data was lost, said Mission Control with Atlantis was restored after eight minutes, but it was an hour before engineers realized crossed signals,**< */t* >

< *h* >**Atlantis is ATLANTIS THE LOST EMPIRE.**< */h* >

< */pair* >

Example2: This is an instance of the AVE test collection that returns an entailment value of 'YES', in this case the LSI Module returns a value of 0.905481:

< *pair* id="7818" value="YES" task="QA">

<q lang="EN" src="clef2006" type="OBJECT"> **What is Atlantis** < */q* >

< *t* doc="LA110794-0104"> **NASA briefly lost contact with the space shuttle Atlantis and its six astronauts Sunday because of crossed radio signals.**< */t* >

< *h* > **Atlantis is the space shuttle.**< */h* >

< */pair* >

The LSI module returns 'YES' if the threshold is equal or above '0.8' and 'NO' for the rest values.

In both examples the LSI module obtains the correct entailment value. In the ***Example1*** the threshold is around 40%, so the answer is 'NO' and in the ***Example2*** the threshold is around 90%, so the answer is 'YES'.

2.3 Module combiner

In the final stage of our AVE system, the previously described word overlap and LSA modules are combined by voting. In order to guarantee that the voting combination is reasonable, we tested the modules for compatibility. One such approach is based on the Kappa coefficient [1], [8], [4] which measures the agreement of the classifiers. High Kappa value corresponds to high agreement between the runs hence no improvement when voting is applied, while low Kappa value corresponds to low agreement and improvement after the combination.

For an answer validation pair, we have obtained different outputs from the LCS, skip-gram and LSA runs. We measured the Kappa agreement for the three runs altogether and also we tested them by pairs. The experiment was carried out for English and Spanish. According to the obtained results, the best combinations are LCS with skip-gram, and LCS, skip-gram, LSA. Therefore, we have submitted two runs for the AVE challenge.

When the Kappa measure determined the outputs that have to be combined, we applied voting. Voting is a technique that aims to combine multiple evidences into a singular prediction. The generated outputs of LCS, skip-gram and LSA are taken and compared. The final decision about the class assignment is taken regarding the class with the majority votes. For the LCS and the skip-gram runs we could not apply voting because the number of classifiers is even. Therefore, we applied the following strategy in which when the two outputs agree, the obtained result remains the same, but when they disagree the AVE pair is assigned with NO value e.g. the answer validation pair is incorrect.

3 Results of the Multilingual AVE Runs

In this section, we present the obtained results for the different languages. We have participated in English, Spanish, German, French, Italian, Dutch and Portuguese. Table 1 shows the results of the individual word overlap sets and the LSA runs as well as the results from the two combinations we have performed. A discussion of the carried out experiments and the obtained results for each one of the languages can be found below.

To evaluate the performance of our system, we have used the following evaluation measures:

$$precision = \frac{\#predicted\ as\ YES\ correctly}{\#predicted\ as\ YES} \tag{1}$$

$$recall = \frac{\#predicted\ as\ YES\ correctly}{\#YES\ pairs} \tag{2}$$

$$f - score = \frac{2 * recall * precision}{recall + precision} \tag{3}$$

These measures are introduced by the AVE organizers, because of the distribution in an AVE corpus. According to their study [7] 25% of the pairs are YES and 75% of them are NO, therefore the performance of an AVE system should be evaluated only with the YES pairs.

English: For this language, we have performed a training phase by merging the ENGARTE data sets provided by the AVE organizers[3]. The obtained results from this experiment served as indicator for the best attributes of the initial feature set. The best word overlap feature both for the train and test data sets is LCS. This shows that one third of the AVE pairs can be resolved correctly simply by considering the overlapping insequences of the words between two texts.
The skip-gram and LSI runs performed also around 26-27%. When the two word overlap attributes are merged there is no improvement for the test data, however in the training phase the increment is 2%. The highest score for this language are obtained when the LCS, skip-gram and LSI runs are merged. This shows that LSI identifies correctly examples which are omitted by the other modules. The voting combination has 8% improvement compared to the performance of a single classifier. According to the z' statistics with confidence level of 0.975, the increment is significant.

Spanish: A separate training phase is conducted for the Spanish language. This time we have trained the word overlap modules with the SPARTE corpus. For the test data, the best score is reached by LCS and has the value of 53.15%. The voting combination of LCS, skip-gram and LSI has the same performance as the individual LCS classifier. This is due to the low coverage of LSI, which depends on the number and type of the words in the T-phrases.

German, French, Italian: For these languages the best performances are obtained with the LCS and voting runs. The obtained f-scores range from 40 to 47%. The performance of LSI is lower than those of the word overlap module, because of the similarity threshold. Although the 0.8 threshold is obtained after we have studied the performance of LSI with the Spanish and English training data, the multilingual test experiments show that the threshold is sensitive to the words of the T-phrases.

[3] http://nlp.uned.es/QA/ave

Dutch, Portuguese: For these two languages, our system obtained the lower scores among all. It is interesting to note that for Dutch the skip-gram run performed better than LCS. This may be related to the origin of the language and the word order, as skip-gram looks for position independent n-grams. For Portuguese, LSI performed better than the word overlap runs. The voting combination for this language has 4% better coverage than the individual classifiers.

Discussion: In this AVE competition, we have participated for the English, Spanish, German, French, Italian, Dutch and Portuguese languages. The carried out experiments show that one and the same attributes can be applied to different languages and even can reach the same performance. Thus, we have proved that our MLEnt system is adaptable to the RTE's subtask AVE and second that our hypothesis for MLEnt's multilinguality is valid.

Globally for Spanish, German and French, the LCS and voting combination performed the same. However, for English the voting combination improved with 8% the performance of the individual classifier. For Italian the voting strategy had an increase of 3% and the same happen with Portuguese. Compared to the individual performance of LSI, the voting combination was better for each one of the seven languages.

The variation of the performances of the word overlap attributes and the combination strategies shows that even in the AVE challenge the voting combination depends on many factors such as sequences of n-grams, number of words in the text, the results of the individual classifiers. Since for most of the languages the voting strategy had positive effect, we can claim that voting improves the performance of our system.

4 Conclusions and Future work

The main aim for our participation in the AVE competition was to test our previously developed text entailment system MLEnt for multilinguality and how it will perform in a new task such as answer validation. As the textual entailment and the answer validation tasks share the same idea of identifying whether two texts infer the same meaning, it was easy and possible to adapt the MLEnt system to the AVE challenge.

With the obtained results we proved that MLEnt can function for different languages. The performance of the system ranges from 20% to 53% depending on the language pairs. It is interesting to note that the performance of our AVE system depends on the LCS, skip-gram and LSI attributes, however the two most robust attributes are LCS and skip-grams which for German, French, Italian and Dutch reached the coverage of 47% for LCS and 38% for skip-gram respectively.

We do not include the Spanish and English results, because we used training data and the performance for these languages is influenced also by the size of the data. The only language with lower performance is Portuguese. Probably because the amount of the overlapping words was not so much as for the rest of the languages.

The performance of the LSI attribute also varied across the languages. As LSI uses the words of the text to construct the conceptual matrices, we observed that for the languages where the T-phrases were longer, the performance of LSI was better and vice versa. As a conclusion from the carried out experiments, we can say that in the future these attributes can serve as a baseline for all languages but should be further improved by the incorporation of richer knowledge such as syntactic or semantic.

During the experiments, the Kappa agreement measure determined the compatible set for the voting combination. By this measure the performance of the individual sets was improved. The voting combination lead to improvement for the English, French, Italian and Portuguese language runs. For Spanish and German the performance of the voting combination and the LCS attribute differed by 0.14%, and according to the z' statistics this difference is insignificant. For Dutch the LCS and skip-gram attributes performed better than the combination. According to the conducted experiments, the attribute which was most informative for the correctness of the answer validation pair is LCS.

The present AVE system has the ability to work fast and is characterized by quick training and testing phases, which is very important for a real Question Answering application. Another

Language runs	Precision	Recall	F-score
English_LCS	15.22	80.93	28.57
English_Skip	16.91	69.30	27.18
English_LSA	23.29	30.23	26.31
English_LCS&Skip	18.33	64.65	28.56
English_LCS&Skip&LSA	24.92	69.77	**36.72**
Spanish_LCS	44.21	66.62	**53.15**
Spanish_Skip	37.24	43.07	39.94
Spanish_LSA	34.15	14.45	20.31
Spanish_LCS&Skip	47.48	39.34	43.03
Spanish_LCS&Skip&LSA	40.65	76.15	**53.01**
German_LCS	38.90	60.56	**47.37**
German_Skip	34.37	43.91	38.55
German_LSA	11.43	1.13	2.06
German_LCS&Skip	41.30	37.68	39.41
German_LCS&Skip&LSA	36.34	67.42	**47.22**
French_LCS	33.96	67.09	**45.09**
French_Skip	30.48	46.38	36.78
French_LSA	32.36	15.88	21.31
French_LCS&Skip	38.36	43.69	40.85
French_LCS&Skip&LSA	34.44	73.62	**46.93**
Italian_LCS	25.78	70.59	**37.77**
Italian_Skip	21.96	86.10	34.99
Italian_LSA	29.16	22.45	25.37
Italian_LCS&Skip	21.64	88.77	34.80
Italian_LCS&Skip&LSA	28.30	72.19	**40.66**
Dutch_LCS	14.26	90.12	24.62
Dutch_Skip	15.80	67.901	**25.64**
Dutch_LSA	13.88	12.34	13.07
Dutch_LCS&Skip	18.90	67.90	**29.57**
Dutch_LCS&Skip&LSA	14.84	90.12	25.48
Portuguese_LCS	12.50	3.90	5.94
Portuguese_Skip	8.00	21.00	11.58
Portuguese_LSA	11.26	12.76	11.96
Portuguese_LCS&Skip	19.04	12.77	**15.29**

Table 1. *Results for the AVE runs*

benefit for our AVE system is coming form the nature of the attributes, which depend only on the length of the sentences and the number of overlapping words. This allowed us to normalize the attributes and to used them as a comparative measure for the languages for which we had no training data.

In the future we want to study the influence of stemming for the different languages. We are also interested in improving the AVE system for Spanish and English, by the incorporation of syntactic information, by the measurement of the similarity of the noun phrases and also by the validation of named entities.

5 Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and PROFIT number FIT-340100-2004-14 and by the Valencia Government under project numbers GV04B-276.

References

1. J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*, 1960.
2. Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the PASCAL workshop on Text Understanding and Mining.*
3. Oren Glickman. *Applied Textual Entailment.* PhD thesis, Bar Ilan University, 2005.
4. Zornitsa Kozareva and Andrés Montoyo. An approach for textual entailment recognition based on stacking and voting. In *Proceedings of 5th Mexican International Conference on Artificial Intelligence, MICAI*, 2006.
5. Zornitsa Kozareva and Andrés Montoyo. Mlent: The machine learning entailment system of the university of alicante. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.
6. T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition. In *Psychological Review*, pages 211–240, 1997.
7. Anselmo Peñas, Alvaro Rodrigo, and Felisa Verdejo. Sparte, a test suite for recognising textual entailment in spanish. In *Proceedings of CICLing*, 2006.
8. Ted Pedersen. Assessing system agreement and instance difficulty in the lexical sample tasks of senseval-2. In USA Philadelphia, PA, editor, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*
9. Sonia Vázquez, Zornitsa Kozareva, and Andrés Montoyo. Textual entailment beyond semantic similarity information. In *Proceedings of 5th Mexican International Conference on Artificial Intelligence, MICAI*, 2006.