SPARTE, a Test Suite for Recognising Textual Entailment in Spanish

Anselmo Peñas, Álvaro Rodrigo, and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo, alvarory, felisa}@lsi.uned.es

Abstract. The aim of Recognising Textual Entailment (RTE) is to determine whether the meaning of a text entails the meaning of another text named hypothesis. RTE systems can be applied to validate the answers of Question Answering (QA) systems. Once the answer to a question is given by the QA system, a hypothesis is built turning the question plus the answer into an affirmative form. If the text (a given document) entails this hypothesis, then the answer is expected to be correct. Thus, a RTE system becomes an Answer Validation system. Within this framework the first problem is to find collections for training and testing RTE systems. We present here the SPARTE corpus aimed at evaluating RTE systems in Spanish. The paper presents the methodology to build SPARTE from the Spanish QA assessments performed at the Cross-Language Evaluation Forum (CLEF) during the last three editions. The paper also describes the test suite and discusses the appropriate evaluation measures together with their baselines.

1 Introduction

The task of Recognising Textual Entailment (RTE) [3] aims at deciding whether the truth of a text entails the truth of another text named hypothesis or, in other words, if the meaning of the hypothesis is enclosed in the meaning of the text. The entailment relation between texts is useful for a variety of tasks as, for example, Automatic Summarisation, where a system could eliminate the passages whose meaning is already entailed by other passages; or Question Answering (QA), where the answer of a question must be entailed by the text that supports the correctness of the answer.

Since RTE task has been defined recently, there exists only few corpora for training and testing RTE systems, and none of them are in Spanish. Thus, we planned the development of SPARTE, a corpus for training and testing RTE systems in Spanish, and specially, systems aimed at validating the correctness of the answers given by QA systems. This automatic Answer Validation would be useful for improving QA systems performance and also for helping humans in the assessment of QA systems output.

SPARTE has been built from the Spanish corpora used at Cross-Language Evaluation Forum (CLEF) for evaluating QA systems during 2003, 2004 and 2005. At the end of development, SPARTE contains 2962 hypothesis with a document label and a TRUE/FALSE value indicating whether the document entails the hypothesis or not.

Section 2 describes the development of SPARTE in detail. Section 3 evaluates some features of the corpus. Section 4 discusses and suggests the way of using SPARTE for evaluation purposes. Section 5 is devoted to some other corpora related to RTE. Finally, we give some conclusions and future work.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 275–286, 2006.

2 Development of SPARTE

SPARTE is a training and testing corpus for RTE systems, containing text and hypothesis pairs together with a TRUE/FALSE value indicating whether the text entails the hypothesis or not. The hypothesis have been built from the questions and answers used in the evaluation of QA systems at CLEF. Next subchapters describes in detail the methodology followed.

2.1 Original Corpus

The starting point for development of SPARTE were the Spanish corpora used at the Cross-Language Evaluation Forum (CLEF) for evaluating Spanish Question Answering (QA) systems during the last three years [7] [9] [8] [5] [11]. The organization provided the participants with the set of questions they had to answer, and a large document collection where the systems had to find and extract the answers. The Spanish collection contains 454,045 news in Spanish from the EFE News Agency for the years 1994 and 1995.

Each year, the participant systems submitted up to two runs responding 200 questions per year. Together with the answer string, the systems must indicate the document that supports the correctness of the answer. Occasionally, systems give NIL to indicate that the question had no answer in the collection. Table 1 gives an idea of the corpus size, showing the number of question and answer pairs available at the beginning of the development. Each answer was assessed by humans in order to decide whether it was correct, exact and supported by the given document or not.

Table 1. Number of question answer pairs for SPARTE development

Year	#questions	#answers	#runs	#participants	#q-a pairs
2003	200	3	2	1	1200
2004	200	1	2	5	1600
pilot 2004	100	-	1	1	100
2005	200	1	2	9	3598
					6498

2.2 Building the Hypothesis

Since textual entailment was defined between statements, the first step was to turn the questions into an affirmative form. For example, the question *"Which is the capital of Croatia?"* was transformed into *"The capital of Croatia is <answer/>"*, where the mark *"<answer/>"* has to be instantiated with any answer given to that question by any system. In this way, we prepared the corpus to build all the hypothesis automatically by substituting the mark with the corresponding answers.

Figure 1 shows the xml format used at this stage. The *answer* mark inside the hypothesis is not instantiated yet. The *instance* marks contain an answer that will substitute the *answer* mark to complete the hypothesis. Instances include the document identification

```
<case id="1">
    <question>
       ¿Cuál es la capital de Croacia?
    </question>
    <hypothesis>
       La capital de Croacia es <answer/>
    </hypothesis>
    <instance id="1" text="EFE19940127-14481" eval="R">
       Zagreb
    </instance>
    <instance id="2" text="EFE19941119-11475" eval="W">
       Fuerzas de la ONU
    </instance>
    <instance id="3" text="EFE19940907-03455" eval="W">
       ONU
    </instance>
    <instance id="4" text="EFE19940907-03366" eval="W">
       Bosnia-Herzegovina
    </instance>
</case>

...

<case id="88">
    <question>
       ¿En qué año Kuwait fue invadido por Irak?
    </question>
    <hypothesis>
       Kuwait fue invadido por Irak en el año <answer/>
    </hypothesis>
    <instance id="1" text="EFE19950202-00912" eval="X">
       2 de agosto de 1990
    </instance>
    <instance id="2" text="EFE19940326-16845" eval="R">
       1990
    </instance>
    <instance id="3" text="EFE19950411-06234" eval="U">
       1990
    </instance>
    <instance id="4" text="EFE19950731-19147" eval="W">
       Bagdad
    </instance>
</case>
```

Fig. 1. XML for the hypothesis templates

and the assessment given to the answer by a human: correct (R), incorrect (W), inexact (X) or unsupported (U).

Repeated answers (instances) and NIL answers have been removed. NIL answers might be correct or not, but in any case, NIL stands for the absence of answer and therefore, there is no answer to validate.

```
<pair id="1" value="TRUE" task="QA">
    <q>
       ¿Cuál es la capital de Croacia?
    </q>
    <t doc="EFE19940127-14481"> </t>
    <h>
       La capital de Croacia es Zagreb
    </h>
</pair>
...
<pair id="614" value="TRUE" task="QA">
    <q>
       ¿Qué torneo ganó Andrei Medvedev?
    </q>
    <t doc="EFE19940424-13985"> </t>
    <h>
       Andrei Medvedev ganó el torneo de Montecarlo
    </h>
</pair>
...
<pair id="26" value="FALSE" task="QA">
    <q>
       ¿Qué país ganó la Copa Davis?
    </q>
    <t doc="EFE19940406-02726"> </t>
    <h>
       Roland Garros ganó la Copa Davis
    </h>
</pair>
...
<pair id="58"value="FALSE" task="QA">
    <q>
       ¿Quién era conocido como el "Zorro del Desierto"?
    </q>
    <t doc="EFE19940205-02731"> </t>
    <h>
       Laguna del Desierto era conocido como el
       "Zorro del Desierto"
    </h>
</pair>
```

Fig. 2. SPARTE corpus excerpt

Notice that no snippet, short passage or sentence is explicitly given for the text, but the identification of a whole document. The reason is that the current assessment at CLEF requests a whole document for supporting the answer. Thus, the answer can be supported by a conjunction of sentences not necessarily in consecutive order inside the document. In other words, verifying the truth of the hypothesis could require more than one inter-related sentences from the document. From our point of view, this approach is realistic, leaving to the RTE system developers the decision of managing the whole text or only a passage extracted previously, containing the answer string.

2.3 Building the Text-Hypothesis Pairs

Once the answers are grouped as possible instances for building the hypothesis, the next step is to build the text-hypothesis pairs with the entailment TRUE/FALSE value. Figure 2 shows the xml format that conforms the final SPARTE structure. The mark <*pair*> includes the TRUE/FALSE value to indicate whether the text entails the hypothesis or not. Inside <*pair*> there are three marks: one containing the original question, a second containing the document identification, and a third containing the hypothesis to be validated with the document.

The hypothesis has been generated automatically from the hypothesis template instantiated with each answer. Thus, some wrong answers could give to the hypothesis not only a wrong semantics but also a wrong syntactic structure (see Figure 3). In our opinion this is a desirable feature for the corpus, allowing the development of syntactic criteria for RTE and also promoting the development of systems robust to some formal and syntactic errors.

```
<pair id="51" value="FALSE" task="QA">
    <q>
       ¿Qué es UNICEF?
    </q>
    <t doc="EFE19950126-152091"'> </t>
    <h>
       UNICEF es China con
    </h>
</pair>
```

Fig. 3. SPARTE sample with syntactic errors

2.4 Determining the Entailment Value

Each text-hypothesis pair in the corpus has associated an entailment value to indicate if the meaning of the hypothesis can be derived from the document. However, the result of the QA assessment was not a binary value and, therefore, a simple mapping is necessary for converting QA assessments into entailment values. We followed this criteria:

– *Correct answer (R)*: the answer to the question is correct and the document support its correctness. Then the text entails the hypothesis and the entailment value is TRUE.

- *Unsupported answer (U)*: although the answer is correct, the document does not permit to affirm the correctness of the answer. The text does not entail the hypothesis and the entailment value is FALSE.
- *Inexact answer (X)*: This is a difficult case also for the human assessors. There is no additional information to decide whether the answer contains too much information or, in the contrary, the answer string is too short to be considered as a correct one. Both cases were tagged as inexact. In the short cases, the resulting hypothesis could be true in the document, although they are not valid answers. In the long cases, although the answer string contains a correct answer the resulting hypothesis might have a different meaning and be false because of the extra information. Thus, we do not have a clear criteria to determine the entailment values in these cases without a human assessment. Fortunately, there are only few cases (6% of the pairs) and we opted for excluding them from the final SPARTE corpus.
- *Incorrect answer (W)*: the answer to the question is wrong. Although the answer could be directly extracted from the text, the joint reformulation of question and answer as a statement (hypothesis) makes very difficult the entailment between text and hypothesis. Therefore, the entailment value in this cases is FALSE. However, we detected few cases were the answer is not responsive but it becomes a hypothesis true in the text. For example, "Japan" was considered a not responsive answer (wrong answer) to the question "Where did explode the first atomic bomb?". However, the resulting hypothesis "The first atomic bomb exploded in Japan" is true in the text. We have studied a 10% of the FALSE pairs in the final corpus finding that a 4% of the hypothesis FALSE (3% of all hypothesis) could be considered TRUE in the text. In other words, although the source answers can be considered not responsive, and in fact the corresponding pairs have been tagged with an entailment value FALSE, they can be found TRUE in the corresponding texts. However, this source of errors is in the same range than the disagreement between the human annotators that made the QA assessments [11].

3 Evaluation of SPARTE

We performed a partial human evaluation of the corpus in order to assess the quality of SPARTE. We took randomly the 10% of TRUE pairs and the 5% of the FALSE ones (see Table 2).

Table 2. Manual evaluation of SPARTE

	Considered	**Correct**	**Incorrect**
Pairs TRUE	70 (10% of TRUEs)	67 (96%)	3 (4%)
Pairs FALSE	113 (5% of FALSEs)	111 (98%)	2 (2%)
Total	183 (6%)	178 (97%)	5 (3%)

We found that errors are in the same range as inter-annotator disagreement in the QA assessments (less than 5%). In fact, the errors we found come from some wrong

```
<pair id="9" value="TRUE" task="QA">
    <q>
       ¿Cuál es el nombre de pila del juez Borsellino?
    </q>
    <t doc="EFE19940718-10595"> </t>
    <h>
       Paolo Borsellino es el nombre de pila del
       juez Borsellino
    </h>
</pair>
...
<pair id="398" value="TRUE" task="QA">
    <q>
       ¿Qué iglesia aprobó los nuevos cánones para la
       ordenación de mujeres?
    </q>
    <t doc="EFE19941202-00867"> </t>
    <h>
       La iglesia Sínodo de la Iglesia Anglicana aprobó los
       nuevos cánones para la ordenación de mujeres
    </h>
</pair>
```

Fig. 4. Sample with an incorrect TRUE

```
<pair id="144" value="FALSE" task="QA">
    <q>
       ¿A cuántos años de prisión fue sentenciado Bettino
       Craxi?
    </q>
    <t doc="EFE19951103-01683"> </t>
    <h>
       Bettino Craxi fue condenado a ocho años de prisión
    </h>
</pair>
```

Fig. 5. Sample with an incorrect FALSE

QA assessments. Some examples are given in Figures 4 and 5. In the first case, incorrect TRUE pairs are due to *Inexact* answers that were judged as *Right*. In the second case, the incorrect FALSE pair is due to a *Right* answer that was incorrectly judged as *Wrong*. We also verified that errors are independant of the entity type requested by the question.

Another interesting feature of SPARTE is that the hypothesis expressions do not appear in the documents. From the 183 pairs studied, only four hypothesis can be found in the text (see one of them in figure 6). This good feature is the result of building the hypothesis as an affirmative expression of the question which conforms a statement not present in the text. However, we found that in 100% of the cases one sentence of the text is enough to support the answer or, in other words, to entail the hypothesis.

```
<pair id="418" value="TRUE" task="QA">
    <q>
       ¿Cuándo ocurrió la catástrofe de Chernobil?
    </q>
    <t doc="EFE19940626-16005"> </t>
    <h>
       La catástrofe de Chernobil ocurrió en abril de 1986
    </h>
</pair>

DOC "EFE19940626-16005" : "... Entre los países que acogerán
a los niños afectados por radiaciones están también Bélgica,
Alemania, Finlandia y Eslovaquia. La catástrofe de Chernobil
ocurrió en abril de 1986. Unas 7.000 personas murieron inme-
diatamente o poco después a causa de las radiaciones, y va-
rios centenares de miles sufren aún sus consecuencias..."
```

Fig. 6. Sample of hypothesis contained in the text

4 Evaluating RTE Systems with SPARTE

The final SPARTE corpus has 2962 text-hypothesis pairs from 635 different questions (4.66 average number of pairs per question). Table 3 shows also the number of pairs with an entailment value equals to TRUE (695) and the number of pairs FALSE (2267).

Table 3. Number of text-hypothesis pairs in SPARTE

	Number	Percentage
Pairs TRUE	695	23%
Pairs FALSE	2267	77%
Total	2962	

The evaluation of a RTE system must quantify its ability to predict the TRUE/FALSE entailment value. Notice that the percentage of pairs FALSE (77%) is much larger than the percentage of pairs TRUE (23%). We decided to keep this proportion since this is the result of real QA systems submission. However, this fact introduces some issues in the evaluation of RTE systems with SPARTE. For example, a baseline RTE system that gives always FALSE would obtain an accuracy equal to 77%, been a very high baseline (see Table 4).

From the RTE evaluation point of view, it would we enough to balance the corpus selecting randomly a 30% of the FALSE pairs and using only these in conjunction with the TRUE pairs. This would yield a corpus big enough and perfectly balanced.

However, from the Answer Validation point of view, a system that validates QA responses does not receive correct and incorrect answers in the same proportion. Thus we think that is useful to mantain the same percentage of TRUE and FALSE pairs that

Table 4. Baselines accuracy in SPARTE

Baselines	Accuracy
Give always TRUE	0.23
Random 50% FALSE 50% TRUE	0.5
Random 77% FALSE 23% TRUE	0.65
Give always FALSE	0.77

Table 5. Baselines for the evaluation proposed

Baselines	Precision TRUEs	Recall TRUEs	F-measure TRUEs
Answer always TRUE	0.23	1	0.37
Random 50% FALSE 50% TRUE	0.5	0.5	0.5
Random 77% FALSE 23% TRUE	0.23	0.23	0.23

systems will receive in an Answer Validation exercise. We think this leads to different development strategies closer to the real exercise that, anyway, must be evaluated with this unbalanced nature.

For this reason, we propose an evaluation based on the detection of pairs with entailment (entailment value equals to TRUE). Turning this into the Answer Validation problem, the proposed evaluation with SPARTE would be focused on the detection of correct answers. This approach has sense since, at the moment, QA systems give more incorrect answers than correct ones, and working towards the detection of the correct ones would be very useful both, for the improvement of systems and for the combination of results coming from different systems.

With this approach we also give a partial solution to the problem of the pairs FALSE that could be considered TRUE in SPARTE (4% of pairs FALSE) (see section 2.4). Although they affect to the training phase, they do not affect to the testing since they are ignored (as they are still FALSE pairs in the corpus).

Therefore, instead of using an overall accuracy as the evaluation measure, we propose to use precision (1), recall (2) and a F measure (3) over pairs with entailment TRUE. In other words, to quantify systems ability to detect the pairs with entailment. Table 5 shows the three baselines with this setting: a system responding always TRUE, a system responding 50% of TRUEs and 50% of FALSEs, and a system responding TRUEs and FALSEs in the same proportion they appear in SPARTE. As shown in the table, the F measure over pairs TRUE becomes a good measure for comparing different systems under the same evaluation conditions.

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

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

$$F = \frac{2 * recall * precision}{recall + precision} \tag{3}$$

5 Related Work

SPARTE corpus is inspired in the corpus used for training and testing RTE systems at the PASCAL RTE Challenge 2004 [3]. This is a corpus available[1] in English containing 567 text-hypothesis pairs for the development and training phase, and 800 pairs for the testing. In these collections the number of pairs with entailment (TRUE) is equal to the number of pairs without entailment (FALSE). All these pairs were selected, filtered and adapted manually from a different number of sources related to different NLP and IR areas such as Information Retrieval, Machine Translation, Information Extraction, Paraphrase Acquisition or Question Answering among others. All the text-hypothesis pairs have a tag indicating the corresponding type of source. The evaluation showed that all systems perform better over some types of tasks than others. Due to the manual effort, this is a general corpus that covers a wide range of linguistic phenomena and makes very difficult the RTE task for automatic systems.

In the first PASCAL RTE Challenge, MITRE decided to cast the RTE problem as one of statistical alignment and for this, they needed a corpus larger than the one provided by the organization [2]. From MITRE point of view, most of the TRUE pairs exhibit a paraphrase relationship in which the hypothesis is a paraphrase of a subset of the text. Therefore, they took a news corpus in which the headline of a news article is often a partial paraphrase of the lead paragraph. After a semi-automatic processing they selected the more promising 100,000 pairs, estimating that 74% of them have an entailment relationship. This corpus lead MITRE to one of the best results in the PASCAL Challenge. Although this is a corpus useful for training statistical RTE systems, the absence of human assessment for every pair impede its use for evaluation purposes.

There are some other works aimed at acquiring paraphrases without the notion of entailment [1] [10] [6]. One corpus available in this direction is the Microsoft Research Paraphrase Corpus[2] [4]. Based on the idea that an event generates hundreds of different news articles in a closed period of time, they decided to apply unsupervised techniques over news clusters for acquiring sentence-level paraphrases. Again, this corpus could be useful for training RTE systems working with English rather than for evaluation purposes.

6 Conclusions and Future Work

SPARTE is the first corpus aimed at evaluating RTE systems in Spanish, containing 2962 text-hypothesis pairs with TRUE/FALSE entailment values. It is specially oriented to the development and evaluation of Answer Validation systems, that is to say, systems aimed at deciding whether the responses of a QA system are correct or not. For this reason, SPARTE deals with the particularities of QA at CLEF: Answers must be supported by documents, and systems give more incorrect than correct answers. We showed here the methodology for developing SPARTE taking advantage of the human assessments made in the evaluation of QA systems at CLEF. We also suggest the appropriate evaluation methodology using SPARTE.

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

[2] Available at http://research.microsoft.com/research/downloads/default.aspx

Future work is oriented to derive some subcollections of SPARTE were TRUE and FALSE pairs appear in the same proportion.

The second research line is to automatically extract the sentence from the document that contains and supports the answer, in order to give it as the text in the entailment pairs.

Finally, we would like to extend this work to the rest of languages used in the QA Track at CLEF, in order to make available a training corpus for an Answer Validation exercise in multiple languages.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the following project: TIC-2003-07158-C04-02, R2D2-SyEMBRA. We are grateful to Víctor Peinado, Jesús Herrera and Valentín Sama of the UNED NLP Group, for the QA assessments.

References

1. R. Barzilay and L. Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23, 2003.
2. J. Burger and L. Ferro. Generating an Entailment Corpus from News Headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, June 2005.
3. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 1–8, April 2005.
4. B. Dolan, C. Quirk, and C. Brockett. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004, Geneva, Switzerland*, 2004.
5. J. Herrera, A. Peñas, and F. Verdejo. Question Answering Pilot Task at CLEF 2004. In C. Peters, J. Gonzalo, M. Kluck, P. Clough, G.J.F. Jones, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images. CLEF 2004.*, volume 3491 of *Lecture Notes in Computer Science*, pages 581–590, 2005.
6. D. Lin and P. Pantel. DIRT Discovery of inference rules from text. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM Press, 2001.
7. B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke. The Multiple Language Question Answering Track at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003.*, volume 3237 of *Lecture Notes in Computer Science*, pages 471–486, 2004.
8. B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images. CLEF 2003.*, volume 3491 of *Lecture Notes in Computer Science*, pages 371–391, 2004.

9. A. Peñas, F. Verdejo, and J. Herrera. Spanish Question Answering Evaluation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2004.*, volume 2945 of *Lecture Notes in Computer Science*, pages 472–483, 2004.
10. Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic Paraphrase Acquisition from News Articles. In *Automatic Paraphrase Acquisition from News Articles. Proceedings of Human Language Technology Conference, San Diego, USA.*, 2002.
11. A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In *Proceedings of CLEF 2005*, 2005.