Textual Entailment at EVALITA 2009

Johan Bos[1], Fabio Massimo Zanzotto[2], and Marco Pennacchiotti[3]

[1] University of Rome "La Sapienza", Italy
[2] University of Rome "Tor Vergata", Italy
[3] Yahoo! Labs, Sunnyvale, CA, United States

Abstract. The first shared task for recognising textual entailment for Italian was organised during the EVALITA 2009 campaign. The task definition followed that of its English counterpart, and consists of the task of when given a pair of texts, determining whether the first entails the second. A corpus of 800 examples pairs (400 for development/training and 400 for testing) was constructed on the basis of Wikipedia revision histories. Two groups of researchers participated, submitting eight runs in total, with a variation in performance ranging from 0.50 (conform a baseline system that always predicts entailment) to 0.71 accuracy.

1 Motivation

The task of determining inferential relations between two portions of text has developed itself as an important benchmark for evaluating natural language processing systems [1]. In particular, the shared task known as Recognising Textual Entailment (RTE), first introduced in 2004 [2], is now in its fifth edition and enjoys a great interest of NLP researchers [3, 4]. The idea of the RTE task — determining whether one text entails another — is also a promising way to assess systems that claim to process meaning, as it is very hard to evaluate meaning representations [5]. Nevertheless, successful systems have to cover all levels of processing well, from low-level tokenisation and morphological analysis, to syntactic and semantic interpretation, whether this is done in a shallow or deep way.

The RTE shared tasks have been organised for systems that process English. In this article we describe the first textual entailment exercise for Italian, as organised by the EVALITA 2009 evaluation campaign. In this evaluation exercise we will focus on the entailment relation between two short Italian texts. In terminology and evaluation method we follow, by and large, the well established RTE exercises for English. All in all we hope that end-to-end NLP systems for Italian benefit from this initiative, and also drives research in Italian NLP to combine existing and develop new components for natural language understanding.

2 Task Definition

A pair of texts consists of T (for text) and H (hypothesis). Textual entailment is defined as a directional relationship between such pairs. Systems have to decide

whether T entails H, which is the case when the meaning of H can be inferred from T within the context induced by T. Another way of thinking about this is whether H contains any new information with respect to T: if it doesn't, H is entailed by T. The hypothesis must be fully entailed by the text. When the inference is very probable (but not completely certain) the entailment relation still holds (i.e., in cases where T is true, H is 'most likely' true as well).

Note that similar referring expressions in text and hypothesis are assumed to have the same referent (including zero-pronouns, that occur frequently in Italian by subject-drop, as in Example 1 below). Different name expansions in text and hypothesis are taken to refer to the same individual or organisation. Furthermore, the use of presupposition of common linguistic and world knowledge is permitted for drawing inferences. For instance, in Example 3 below, the fact that Napolitano is the Italian President is a commonly known fact, and therefore the entailment relation holds for this pair.

Example 1: entailed

T: *Parla di attività nei panni di direttore commerciale e, dopo sei mesi, di direttore generale.*
H: *Parla di attività di direttore commerciale e, dopo sei mesi, di direttore generale.*

Example 2: not entailed

T: *Il primo acquisto immobiliare fu un terreno in via Alciati a Milano, per 190 milioni di lire.*
H: *Il primo acquisto è un terreno in via Alciati a Milano.*

Example 3: entailed

T: *Napolitano non ha apposto la firma sul decreto.*
H: *Il Presidente non ha apposto la firma sul decreto.*

The development and test data each consist of 400 examples of such pairs, equally divided into positive and negative examples. Performance is measured by accuracy (number of correctly recognised pairs divided by the total number of pairs), thereby setting the baseline score at 50% for a coin-flipping approach. Texts are generally short, covering a sentence. The development and test data was provided to the participants in XML marked-up files, using UTF-8 encoding for Unicode (Figure 1).

3 Dataset Description and Development

The development and test data will consists of completely new annotated data (because it is the first time that the RTE challenge is organised for Italian).

```
<?xml version='1.0' encoding='UTF-8'?>
<entailment-corpus>
<pair entailment="YES" id="0001" task="WIKI">
   <t>Sposato con due figli, diplomato all'Istituto per Geometri
      'Giacomo Quarenghi' di Bergamo, è funzionario al Genio
      civile di Bergamo.</t>
   <h>Sposato, due figli, è funzionario al Genio civile di
      Bergamo.</h>
</pair>
<pair entailment="NO" id="0002" task="WIKI">
   <t>Alle elezioni politiche 2006 è candidato al Senato, ma non
      risulta eletto.</t>
   <h>Dopo le elezioni politiche 2006 conferma il suo seggio a
      Montecitorio, risultando vincitore nella circoscrizione Calabria
      con la lista dell'L'Ulivo.</h>
</pair>
</entailment-corpus>
```

Fig. 1. Snapshot of the XML-encoded entailment corpus.

Pairs of texts will be taken from Italian Wikipedia articles, and are constructed by manually annotating contrasting texts taken from the version history as provided by Wikipedia. In this section we will discuss why we think Wikipedia is a good candidate for giving relevant sentence pairs, and how we performed the extraction of T-H pairs.

3.1 Extracting RTE-pairs from Wikipedia Revisions

Wikipedia is an open encyclopedia, where any person can behave as an author, inserting new entries or modifying existing ones. Our main intuition in using Wikipedia to build an entailment corpus is that a wiki-based framework, with its built-in revision system, should provide a natural source of non-artificial examples comprising true and false entailments.

The primary concern of Wikipedia authors is to reshape a document according to their intent, by adding or replacing pieces of text. Excluding vandalism, there are several reasons for making a revision: missing information, misspelling, syntactic errors, and, more importantly, disagreement on the content. Let's call an original entry S_1 a piece of text in Wikipedia before it is modified by an author, and a revision S_2 the modified text. Then, for example, in Fig. 2, S_1'' is revised to S_2'', as the author disagrees on the content of S_1''. This suggest that when performing an revision, authors concentrate on no other task but revisioning.

Our hypothesis is that (S_1, S_2) pairs represent good candidates of both true and false entailment pairs (T, H), because they represent semantically close pieces of texts. Moreover, (S_1, S_2) pairs are not artificially constructed, as we extract them from pieces of original texts, without any significant modification or post-processing. Also, we believe that pairs extracted from Wikipedia cover

a large range different types of entailment relations, whose distribution is arguably a reliable sample of language in use. In fact, it has been shown that web documents (such as those from Wikipedia) are reliable samples of language [6].

Finally, the Wikipedia texts are not biased in lexical overlap: A sentence S_2 replacing S_1, usually changes only a few words. Yet, the meaning of S_2 may or may not change with respect to the meaning of S_1. In other words, the lexical overlap of the two sentences is very high, but the entailment relation between S_1 and S_2 may be either positive or negative. For example, in Fig. 2 both pairs have high lexical overlap, but the first is a positive entailment ($S'_1 \rightarrow S'_2$), while the second is negative ($S''_1 \rightarrow S''_2$).

S'_1	*Tutt'ora, nel 2008, e a 40 anni d'etá, delizia seppur con qualche pausa a causa di qualche infortunio in piú, i suoi tifosi.*
S'_2	*Tutt'ora, nel 2008, e a 40 anni d'etá, delizia seppur con qualche infortunio in piú, i suoi tifosi.*

S''_1	*In carcere si convertí al cattolicesimo, si sposó e visse fino al 1981, senza che di lei si sapesse mediaticamente piú nulla.*
S''_2	*In carcere si convertí al cattolicesimo, si sposó e visse fino al 1981, senza che di lei si sapesse piú nulla.*

Fig. 2. Sentence pairs from the Wikipedia revision corpus.

3.2 Annotation Guidelines

A raw set of sentence pairs extracted from Wikipedia needs to be human annotated in order to classify the different pairs as positive, negatives and invalid pairs. The authors of this article annotated a sample of several thousand (S_1, S_2) pairs extracted randomly from the Italian Wikipedia. The annotators classified each pair into one of the following classes:

1. **bidirectional**: S_1 entails S_2 and viceversa ($S_1 \leftrightarrow S_2$);
2. **left**: S_1 entails S_2, but not viceversa ($S_1 \rightarrow S_2$);
3. **right**: S_2 entails S_1, but not viceversa ($S_2 \rightarrow S_1$);
4. **no**: neither S_1 entails S_2, nor viceversa ($S_1 \neq S_2$);
5. **reject**: rejected pairs (see guidelines below).

The promote inter-annotator agreement and a consistent dataset, a set of annotation guidelines was developed, which mostly follow those used for building other RTE corpora [4]. We enriched the general RTE guidelines with a few specific ones, related to Wikipedia revisions. In general, the annotators were asked to classify sentence pairs subject to vandalism and extragrammatical text as **reject**.

Embedded Text In many cases, one text (either S_1 or S_2) is composed by more than one sentence, where one of the sentences exactly corresponds to the other text (i.e. is embedded). Such cases should be classified as **reject**. Example:

S_1: Il 5 giugno 2009 viene ufficialmente sostituito come allenatore del Palermo da Walter Zenga; il 12 giugno ha poi dichiarato di aver risolto il contratto con la societ di Viale del Fante.

S_2: Il 5 giugno 2009 viene ufficialmente sostituito come allenatore del Palermo da Walter Zenga; il 12 giugno ha poi dichiarato di aver risolto il contratto con la societ di Viale del Fante. Il 15 giugno 2009 viene nominato ufficialmente nuovo allenatore della SS Lazio, firmando un contratto biennale da 750.000 Euro a stagione.

Slightly Modified Embedded Text These are typical cases in which the author of S_2 adds more information with respect to S_1 by introducing one or more new sentences. From an RTE perspective, these cannot be intended as relevant entailment examples, as textual inference does not play any role at the sentence level. We then ask the annotator to reject such cases. In other cases, a text contains an embedded sentence that does not exactly corresponds to the other text, but is a paraphrase. In such cases we say that the former text entails the latter (i.e. we classify it either as left or right entailment).

Intersentential Anaphoric Reference When either S_1 or S_2 contains a inter-textual pronoun and the other text doesn't we classify the example as *reject*, as in the following:

S_1: Chiusa la carriera a Palermo, segue di nuovo Silvio Baldini questa volta a Parma Calcio e con una nuova veste, quella di allenatore in seconda.

S_2: Chiusa la carriera a Palermo, Atzori segue di nuovo Silvio Baldini questa volta a Parma Calcio e con una nuova veste, quella di allenatore in seconda.

3.3 Final Corpus description

We selected specific domains of the Wikipedia documents for building the entailment corpus (Table 1). In particular, we selected domains of which we expected to be likely to find entries where authors with different views slightly change

Table 1. Distribution of Wikipedia topics in training and testing.

Target	Domain	Entries	Pairs
Training	Italian, German, Cuban, US, and Russian Revolutionary People; Polititians of Popolo della Libertà, Partito Democratico, and Forza Italia; Massons	881	7,651
Testing	Soccer Players (a–d)	1,198	7,720
	Soccer Referees	86	
	Important Cases	62	

the content in order to impose their own opinion on the topic. For this reason, we picked polititians and revolutionary people for selecting the training pairs. Here, even mature documents (i.e., documents that don't grow) slightly change in order to take into account different points of views. The same applies to the domains that we used for extracting the testing pairs. Here we selected soccer players and referees — both are very hot topics in the Italian Wikipedia. We also selected important court cases such as the Enzo Tortora's case.

All in all, we analyzed 881 entries for the training data and 1,346 entries for the testing part. For each entry, we collected the last 50 revisions. We then analyzed two contiguous revisions and we selected only pairs of text that changed. We ended up with 7,651 pairs for training generation and 7,720 pairs for testing generation. After annotation, 400 pairs made it to final training and 400 to the testing examples, respectively.

4 Participation Results

Two participants took part in the first shared task on Italian textual entailment: a team from FBK Irst (Trento, Italy) and a joined team from the University of Alicante (Spain) and the University of Pisa (Italy). Each of these teams submitted four runs. The performance of these runs on the test data is shown in Table 2. When interpreting these results, recall that a baseline system, predicting entailment for each example pair, would yield an accuracy of 50%.

Table 2. Results of all submitted runs for 400 pairs, sorted on accuracy.

Run	Correct	Accuracy
FBKirst_run1.txt	285	0.71
FBKirst_run2.txt	282	0.71
ofe_semTypes_1.txt	257	0.64
ofe_semTypes_2.txt	228	0.57
ofe_lexical_2.txt	230	0.58
ofe_lexical_1.txt	225	0.56
FBKirst_run4.txt	202	0.51
FBKirst_run3.txt	199	0.50

The system from Alicante/Pisa employed a machine learning classifier fed by features derived from lexical distances, part-of-speech information and semantic knowledge taken from SIMPLE-CLIPS, an Italian language resource. This system obtained 58% accuracy when only lexical features were selected. By considering also semantic knowledge, accuracy reached up to 64% [7].

The best run from FBK Irst's system, EDITS (Edit Distance Textual Entailment Suite), a freely available open source tool for Recognizing Textual Entailment (RTE), performed with a 71% accuracy, the highest ranking score of all eight submitted runs [8]. This system achieved best with a token edit distance

algorithm, while the use of deep syntactic analysis did not improve results, an observation that Cabrio *et al.* attribute to the high word overlap and shared similar syntactic structures between the sentence pairs.

5 Discussion

Compared to the English edition, the Italian RTE shared task attracted relatively few participants. It is probably fair to say that this isn't completely unsurprising, as most of the research on Natural Language Processing focusses on English. Yet the question arises whether it is too early in the development of Italian NLP for organising a shared task that requires complete systems comprising several layers of linguistic analysis and having access to lexical resources.

References

1. Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Pinkal, M., Milward, D., Poesio, M., Pulman, S.: Using the Framework. Technical report, FraCaS: A Framework for Computational Semantics (1996) FraCaS deliverable D16.
2. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Lecture Notes in Computer Science. Volume 3944. (2006) 177–190
3. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D.: The second pascal recognising textual entailment challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy (2006) 1–9
4. Sekine, S., Inui, K., Dagan, I., Dolan, B., Giampiccolo, D., Magnini, B., eds.: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, Prague (June 2007)
5. Bos, J.: Let's not argue about semantics. In: Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco (2008) 2835–2840
6. Keller, F., Lapata, M.: Using the web to obtain frequencies for unseen bigrams. Computational Linguistics **29**(3) (2003) 459–484
7. Ferrández, O., Toral, A., Muñoz, R.: Exploiting Lexical Measures and a Semantic LR to Tackle Textual Entailment in Italian. In: Proceedings of EVALITA 2009, Reggio Emilia (2009)
8. Cabrio, E., Mehdad, Y., Negri, Matteo Kouylekov, M., Magnini, B.: Recognizing Textual Entailment for Italian, EDITS @ EVALITA 2009. In: Proceedings of EVALITA 2009, Reggio Emilia (2009)