Techniques for Recognizing Textual Entailment and Semantic Equivalence

Jesús Herrera, Anselmo Peñas, and Felisa Verdejo

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Madrid, Spain
{jesus.herrera, anselmo, felisa}@lsi.uned.es

Abstract. After defining what is understood by textual entailment and semantic equivalence, the present state and the desirable future of the systems aimed at recognizing them is shown. A compilation of the currently implemented techniques in the main Recognizing Textual Entailment and Semantic Equivalence systems is given.

1 Introduction

The concept "textual entailment" is used to indicate the state in which the semantics of a natural language written text can be inferred from the semantics of another one. More specifically, if the truth of an enunciation entails the truth of another enunciation. For example, given the texts:

1. The three-day G8 meeting will take place in Scotland.
2. The Group of Eight summit will last three days.

It is clear that the semantics of the second one can be inferred from the semantics of the first one; then, it is said that textual entailment exists between both texts. Textual entailment is a directional relationship: in the example above, the first statement entails the second one, but this entailment is not given in the opposite direction. The recognition of textual entailment requires a processing at the lexical level (for example, synonymy between *meeting* and *summit* or between *G8* and *Group of Eight*), as well as at the syntactic level and the sentence semantic level. Entailment between natural language texts has been studied in the last years, either as a part of more complex systems or as an independent application. The long-term interest for Recognizing Textual Entailment (RTE) systems is to give service to a wide range of applications which need to determine entailments between pieces of text written in natural language.

When the entailment relation is verified in both directions, then there is a semantic equivalence between the pair of statements, sometimes named paraphrase. Lin and Pantel [17] show a classification of the fields in which the recognition of semantic equivalence is useful, identifying the following:

– Language generation: where efforts in the detection of semantic equivalence have been focused mainly on rule-based text transformations, in order to satisfy external restrictions such as length and readability.

R. Marín et al. (Eds.): CAEPIA 2005, LNAI 4177, pp. 419–428, 2006.

- Automatic summarization: in which the detection of paraphrasing is relevant for avoiding redundancy between statements in a summary.
- Information Retrieval: in which is common to identify phrasal terms from queries and to generate their variants for query expansion.
- Text mining: in which a goal is to find semantic association rules between terms.

Other applications, such as answer validation in question answering tasks or translation comparison in machine translation, can be added.

Classically, the detection of entailment between texts has been tackled by means of some kind of calculus applied to abstract representations of texts. The formalism used to represent texts depends on the kind of treatment given to them. When applying a surface treatment to texts, they can be represented by means of formalisms such as syntactic trees. But when a deep treatment is accomplished, more complex formalisms are needed in which different normalization levels can be given; for example, a normalization level between active voice and passive voice, or a deeper normalization level, as the one proposed by Schwank, using primitives. Thus, the richness of every inference level varies with the kind of normalization level. The kind of calculus necessary to determine when a pair of texts hold entailment depends on the representation formalism selected. Therefore, when using representations corresponding to a surface treatment of the texts, usually similarity metrics between representations are computed; but when using a deep treatment, logic calculus, theorem provers, etcetera are the more suitable techniques in order to detect entailment. Apart from these classic techniques of Natural Language Processing, the advent of mass access to textual information in digital format has meant the success for empirical methods, such as statistical analysis and machine learning. These methods are usually applied in a quite superficial level of knowledge representation.

Despite the number of systems aimed at determining the existence of equivalence and entailment relations between pieces of text written in natural language, there is not a systematization of techniques and tools for the development of such kind of systems. In the following sections, a compilation of the currently implemented techniques in the main Recognizing Textual Entailment and Semantic Equivalence systems is given.

2 Linguistic Techniques

One way or another, all the techniques for linguistic processing are liable of being included in a RTE or a Semantic Equivalence based system. Following, the used ones for developing this kinds of systems are shown:

2.1 Preprocessing

Apart from the necessary token identification, there are systems that develop a preprocessing of the texts before applying them a morphosyntactic analysis, which is stated at the bottom of linguistic processing levels. This processing

consist, in most cases, in the segmentation of sentences and phrases, which has been used as a preparation for the morphological analysis or for the creation of structures for representing texts.

The MITRE[1] is an example for it. They apply to the texts and the hypothesises a sentence segmenter, previously to the morphological analysis.

The system of the Concordia University [2] does not accomplish a morphological analysis but a noun phrase chunking as a basis for creating predicate structures with arguments for every text and hypothesis. A similarity metric between the structures of every pair of text snippets is established in order to determine if there is an entailment between them.

2.2 Morphological and Lexical Analysis

From this kind of analysis, they can be distinguished the following cases: lemma or stem extraction, part-of-speech tagging, use of morphological analyzers and extraction of relations forced by derivational morphology.

The morphological analysis has been used as a first text processing in order to obtain information for subsequent stages which permit to assess the entailment between texts.

The **lemma extraction** is a fairly profusely used technique and, in some cases, it supposes a great part of the total processing accomplished by RTE systems. Lemmatization is necessary not only for accessing lexical resources as dictionaries, lexicons or *wordnets* but it has been used with three different goals: to assess the coincidence between lemmas in similarity measures when treating texts as bags of words, as attributes of graph representations of the texts, and to fit parameters in assessing similarity algorithms. Therefore, as an example, the universities of Edinburgh and Leeds' system [5] uses lemmatization as the most sophisticated language processing; after it, only an overlap measure between lemmas from the hypothesis and the text is applied to determine the existence of an entailment between them. The University of Illinois at Urbana-Champaign' system [8] uses lemmas as a part of the attributes associated to the nodes of concept trees which represent both the texts and the hypothesises. The University of Rome "Tor Vergata" and the University of Milano-Bicocca [19] developed a system in which a morphological analysis is applied for lemma extraction; these lemmas are used in combination with tokens and other items for fitting – by means of a SVM[2] learning algorithm – the parameters of an overall similarity measure between the two graphs representing the text and the hypothesis.

The **stem extraction** has been a technique basically used to obtain data as an input for other system's modules. The use of stems in monolingual English is justified because the good performance shown, motivated by the simplicity of the English morphology; in the future, when RTE systems will be developed for other languages, it will be necessary to assess the possibility of working only with stems or, on the contrary, it will be compulsory to use lemmas. As an example,

[1] The MITRE Corporation, United States.
[2] Support Vector Machine.

the system of the universities "Tor Vergata" and Milano-Bicocca [19] compares stems in order to measure the subsumption of nodes of the graphs they use to represent textual information. This measure, in conjunction with other measure for the subsumption of edges, determine the overall subsumption between graphs representing the text and the hypothesis; the overall subsumption measure is useful to detect the entailment between the text and the hypothesis.

The **part of speech tagging** has been used in two different ways: the system of the MITRE [4] and the one of the University Ca' Foscari and the ITC-irst[3] [9] include it as a linguistic analysis module in a typical cascaded system; but the University of Illinois at Urbana-Champaign [8] uses parts of speech as a subset of the attributes associated to the nodes of conceptual trees representing both the text and the hypothesis.

The **use of morphological analyzers** as such was accomplished only by the MITRE [4], applying a morphological analyzer (Minnon et al., 2001) which action was added to the part of speech tagging, and the results were used as an input for the following stages (a syntactic analyzer of syntactic constituents, a dependency analyzer and a logic proposition generator).

The **extraction of relations given by derivational morphology** is a not frequently used technique; an example can be found in the system of the Language Computer Corporation [10], which extracts relations between words from *WordNet* derivational morphology.

2.3 Multiword Recognition

Is a not widely used technique. For example, the system of the UNED[4] [13] uses it to detect entailment between lexical units; for this, a fuzzy search of the multiwords of the texts in *WordNet* is accomplished, by means of the Levenshtein distance. It permits to establish semantic relations (synonymy, hyponymy, etcetera) not only between words but between multiwords and words.

2.4 Numerical Expressions, Temporal Expressions and Entity Recognition

There are not very used techniques yet. In the case of entity recognition, two examples can be found only: the Stanford University [20] and the University of Illinois at Urbana-Champaign [8]. Stanford's system detects named entities and resolves correferences, aiming at finding dependencies between nodes of the graphs representing the texts. The one of the University of Illinois, uses named entities as attributes of the nodes of the graphs representing the texts. As for the detection of numeric and temporal expressions, two other examples can be found: the Stanford University [20] accomplishes a treatment of numeric expressions, being able to determine inferences like *"2113 is more than 2000"*. The

[3] ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Scientific and Technological Research Center, Italy.

[4] Universidad Nacional de Educación a Distancia, Spanish Distance Learning University.

University Ca' Foscari and the ITC-irst [9] detect temporal expressions, in order to accomplish coherence checks.

2.5 Syntactic Analysis

The **dependency analysis** is one of the most used techniques; probably, this situation has been favored by the public availability of dependency analyzers for the English language showing a high efficiency and a high recall, such as the one developed by Dekang Lin [16] (Minipar).

Using Minipar, Dekang Lin and Patrick Pantel [17] proposed a non-supervised method for the extraction of inference rules from text (DIRT algorithm); some examples of these rules are the following: *"X is author of Y" = "X wrote Y"* , *"X solved Y" = "X found a solution to Y"*, *"X caused Y" = "Y is triggered by X"*. Their algorithm is based on an extended version of the Harris' Distributional Hypothesis [12], which states that words that occurred in the same contexts tend to be similar; instead of it, they applied the hypothesis not to words but paths from dependency trees obtained from a corpus of texts. Lin and Pantel's work aimed at simplifying the creation of knowledge bases for this kind of rules, which usually is done manually and it is very laborious.

In most cases, a parsing tree representing the analyzed text is obtained; but it is used as an auxiliary to obtain a logic representation, too. Examples for the former kind of use are UNED's system [13] and the team of the University of Trento and the ITC-irst's system [15]. The first one assesses the existence of entailment between text and hypothesis by means of the overlap between the dependency trees of both text snippets. The second one assesses the existence of entailment between text and hypothesis by means of the editing distance between the dependency trees of both text snippets; it is based on the previous work of Hristo Tanev, Milen Kouylekov and Bernardo Magnini [21], who developed a textual entailment recognizing system in order to use it as a subsystem of a question answering system. As an example for the other kind of use, the MITRE [4] implements a set of cascaded linguistic analysis subsystems, which includes a stage for dependency analysis; before the dependency analyzer there is a constituent syntactic analyzer, and a logic predicate generator after it.

The **constituent analysis**, on the other hand, is a not very used technique. The University "Tor Vergata" and the University of Milano-Bicocca [19] use constituents in order to extend dependency graphs. The University Ca' Foscari, with the ITC-irst [9], accomplish a constituent analysis as a part of a hybrid syntactic analysis.

2.6 Semantic Analysis

The **semantic roles** tagging was used by the universities of Illinois at Urbana-Champaign [8], Stanford [20] and Ca' Foscari in association with the ITC-irst [9]. The system of Illinois at Urbana-Champaign searches for coincidences between sets of attributes and the structure of the arguments, either at the semantic role level either at the syntactic analysis level. For the case of Stanford, this tagging

permits to add relations between words not previously identified by means of the syntactic analysis; in addition, it permits to classify temporal and locative sentences. In all these cases, the tags were applied to the nodes of the graphs representing text snippets. The University Ca' Foscari and the ITC-irst used semantic roles in a similarity measure between the text and the hypothesis, by means of the count of similar tags between the ones of the text and the ones of the hypothesis.

Some systems represent texts in a logic form after a linguistic analysis, such as the one of the University Macquarie [1]. This one uses an **automated deduction** system that compares the atomic propositions obtained from the text and the hypothesis in order to determine the existence of entailment.

3 Other Techniques

Apart from the techniques showed before, a significant part of the systems implement one or more of the following:

3.1 Using Thesauri, Big Corpora and *WordNet*

An important part of the systems obtains knowledge from thesauri, big corpora and *WordNet*. The queries to *WordNet* have been launched either searching for the acquisition of relations between lexical units from the relations of *WordNet* – such as UNED's system, which searches for synonymy, hyperonymy and *WordNet* entailment relations in order to detect entailments between lexical units from the text and the hypothesis –, either for the obtention of relations from lexical chains, such as University of Concordia's system [2]. Thesauri have been used in order to extract knowledge of concrete fields such as geographical knowledge, obtained by the universities of Edinburgh and Leeds [5] from the "CIA factbook". Big corpora such as the web or the Gigaword newswire corpus have been used in order to acquire lexical properties [4] or co-occurrence statistics [11].

3.2 Paraphrase Detection

The use of paraphrases aims at the obtention of rewriting rules, in order to improve performance when determining if two expressions are equivalent or not. As an example, the University of Illinois at Urbana-Champaign, from a paraphrasing rules corpus developed by Lin y Pantel (2001), obtained a set of rewriting rules, which were used by their system in order to generate variants of the texts [8].

3.3 Machine Learning

Some systems used this kind of algorithms such as, for example, the one of the universities 'Tor Vergata" and Milano-Bicocca [19], which applied a SVM algorithm in order to assess the parameters of an evaluation measure.

3.4 Definition of a Probabilistic Frame

The only existing example is the University Bar Ilan's one, which defines a probabilistic frame in order to modelize the notion of textual entailment [11]; in addition, it uses a bag of words representation in order to describe a lexical entailment model from co-occurrence statistics obtained from the web. It is said that a text probabilistically entails a hypothesis if the text increases the likelihood of the hypothesis being true. In order to treat lexical entailment, a probabilistic model is established for which a word of the hypothesis must be entailed by other word of the text, in a similar way as done in statistical machine translation [6]. Therefore, the probabilistic entailment between text and hypothesis is computed according to the referred lexical entailment. The probabilities of lexical entailment are empirically estimated by means of a non-supervised process based on web co-occurrences.

3.5 Machine Translation

The MITRE developed a system inspired in statistical machine translation models [4] that: trains a machine translation system by means of leads and headlines from a newswire corpus; estimates manually the reliability of the previous training; trains a text classifier for refining the previously obtained corpus; inducts aligning models from the selected subset of the newswire corpus; combines all the features using a k-nearest-neighbour classifier that chose, for every pair <text, hypothesis>, the dominant truth value among the five nearest neighbours in the development set.

4 Evaluation and Corpora

The First PASCAL[5] RTE Challenge [7], aimed at providing an opportunity to present and compare diverse approaches for modeling and for recognizing textual entailment. The task that systems had to tackle was the automatic detection of semantic entailment between pairs of texts written in natural language (monolingual English). For this purpose, the organizers provided to the participants two corpora, one for training and one for testing. The corpora were conformed by pairs of short texts in natural language pertaining to the press news domain. The components of a pair were named as "text" and "hypothesis", respectively. The systems had to detect if the meaning of the hypothesis could be inferred from the meaning of the text. The pairs <text, hypothesis> conforming the corpora provided to the participants of the PASCAL RTE Challenge were chosen so that typical features of diverse text processing applications were present; therefore, the following classification was obtained: Information Retrieval, Comparable Documents, Reading Comprehension, Question Answering, Information Extraction, Machine Translation and Paraphrase Acquisition. The Second PASCAL RTE Challenge [3] toke place while this paper was been revised. It was

[5] Pattern Analysis, Statistical Modeling and Computational Learning. `http://www.pascal-network.org/`.

very similar to the First Challenge, but the tasks considered this time for the classification of the pairs were: Information Extraction, Information Retrieval, Multi-Document Summarization and Question Answering.

Between the two PASCAL RTE Challenges, some other actions related to RTE and semantic equivalence have been performed. In the ACL[6] Workshop on Empirical Modeling of Semantic Equivalence and Entailment, several items about how to analyse and to develop the kinds of systems of interest,s and how to build corpora for training and testing them were treated, following the ideas given in the First PASCAL RTE Challenge. Related to RTE in Spanish, two initiatives were accomplished by the UNED NLP Group[7]: *a)* the development of the SPARTE test suite for Spanish[18], which is based on the answers given by several systems in Question Answering (QA) exercises from the CLEF[8], and *b)* the organization of an Answer Validation Exercise[9] in order to apply RTE systems for emulating human assessment of QA responses and decide whether an answer is correct or not according to a given text snippet.

5 Conclusions

Broadly speaking, just after the First PASCAL RTE Challenge some tendencies in the development of RTE systems can be distinguished: *a)* Those treating texts as bags of words, being lemma extraction the deeper linguistic analysis accomplished. *b)* Those based on a syntactic representation of texts, including some morphological and lexical processings in order to increase system's performance; in this case, overlap between dependency trees is the preferred technique. *c)* Those accomplishing a deep linguistic treatment, by means of a classical cascaded analysis, covering a wide range of levels: morphological, lexical, syntactic and semantic.

There are little examples of systems implementing only statistical treatments or systems accomplishing a deep linguistic analysis.

The results obtained in the First PASCAL RTE Challenge are not significant about the suitability of the used techniques, because all the participants achieved very similar values of accuracy, ranging between 49.5 % and 58.6 % [7]. But the results of the Second Challenge permit to glimpse what are the more suitable techniques to tackle the Recognizing Textual Entailment problem: while most of the systems ranged between 50.9 % and 62.6 % accuracy – showing a remarkable overcome with respect to the previous Challenge's results – two teams from the Language Computer Corporation reached 73.8 % and 75.4 % accuracy, respectively [3]. One of these latter systems (73.8 % accuracy) exploits the logical entailment between deep semantics and syntax of texts and hypothesises as well as shallow lexical alignment of the two texts [22]; the other one (75.4 % accuracy)

[6] The Association for Computational Linguistics (USA). `http://www.aclweb.org/`.
[7] Natural Language Processing and Information Retrieval Group at the Spanish Distance Learning University. `http://nlp.uned.es/`.
[8] Cross Language Evaluation Forum. `http://www.clef-campaign.org/`.
[9] `http://nlp.uned.es/QA/AVE/`.

utilizes a classification-based approach to combine lexico-semantic information derived from text processing applications with a large collection of paraphrases acquired automatically from the web [14].

Some tasks using RTE are arising and, hopefully, more new tasks will be launched in the near future. These tasks could determine the way in which RTE systems will be developed.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology. Project TIC-2003-07158-C04-02: R2D2-SyEMBRA.

References

1. E. Akhmatova. Textual Entailment Resolution via Atomic Propositions. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 61–64, April 2005.
2. A. Andreevskaia, Z. Li, and S. Bergler. Can Shallow Predicate Argument Structures Determine Entailment? In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 45–48, April 2005.
3. R. Bar-Haim, I. Dagan, B. Dollan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy*, pages 1–9, April 2006.
4. S. Bayer, J. Burger, L. Ferro, J. Henderson, and A. Yeh. MITRE's Submissions to EU PASCAL RTE Challenge. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 41–44, April 2005.
5. J. Bos and K. Markert. Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 65–68, April 2005.
6. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation. In *Computational Linguistics 19(2)*, 1993.
7. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 1–8, April 2005.
8. R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons. Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 29–32, April 2005.
9. R. Delmonte, S. Tonelli, M. A. Picollino Boniforti, A. Brsitot, and E. Pianta. VENSES – a Linguistically-Based System for Semantic Evaluation. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 49–52, April 2005.
10. A. Fowler, B. Hauser, D. Hodges, I. Niles, A. Novischi, and J. Stephan. Applying COGEX to Recognize Textual Entailment. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 69–72, April 2005.

11. O. Glickman, I. Dagan, and M. Koppel. Web Based Textual Entailment. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 33–36, April 2005.
12. Z. Harris. Distributional Structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–37, 1985.
13. J. Herrera, A. Peñas, and F. Verdejo. Textual Entailment Recognition Based on Dependency Analysis and WordNet. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 21–24, April 2005.
14. A. Hickl, J. Williams, J. Bensley, K.Roberts, B. Rink, and Y. Shi. Recognizing Textual Entailment with LCC's GROUNDHOG System. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy*, pages 80–85, April 2006.
15. M. Kouylekov and B. Magnini. Recognizing Textual Entailment with Tree Edit Distance Algorithms. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 17–20, April 2005.
16. D. Lin. Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems, Granada, Spain*, May 1998.
17. D. Lin and P. Pantel. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328, 2001.
18. A. Peñas, Á. Rodrigo, and F. Verdejo. Sparte, a test suite for recognising textual entailment in spanish. In *CICLing*, pages 275–286. Springer, 2006.
19. M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto. Textual Entailment as Syntactic Graph Distance: a Rule Based and a SVM Based Approach. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 25–28, April 2005.
20. R. Raina, A. Haghighi, C. Cox, J. Finkel, J. Michels, K. Toutanova, B. MacCartney, M. C. de Marneffe, C. D. Manning, and A. Y. Ng. Robust Textual Inference using Diverse Knowledge Sources. In *Proceedings of the First PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK*, pages 57–60, April 2005.
21. H. Tanev, M. Kouylekov, and B. Magnini. Combining Linguistic Processing and Web Mining for Question Answering. In *Proceedings of the 2004 Edition of the Text Retrieval Conference*, 2004.
22. M. Tatu, B. Iles, J. Slavik, A. Novischi, and D. Moldovan. COGEX at the Second Recognizing Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy*, pages 104–109, April 2006.