Towards Component-Based Textual Entailment

Elena Cabrio[1,2] and Bernardo Magnini[1]
[1]FBK-irst, Trento, Italy
[2]University of Trento, Italy
{cabrio,magnini}@fbk.eu

Abstract

In the Textual Entailment community, a shared effort towards a deeper understanding of the core phenomena involved in textual inference is recently arose. To analyse how the common intuition that decomposing TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint, we propose a definition for *strong component-based TE*, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. We review the literature according to our definition, trying to position relevant work as more or less close to our idea of strong component-based TE. Several dimensions of the problem are discussed: *i)* the implementation of system components to address specific inference types, *ii)* the analysis of the phenomena relevant to component-based TE, and *iii)* the development of evaluation methodologies to assess TE systems capabilities to address single phenomena in a pair.

1 Introduction

The Recognizing Textual Entailment (RTE) task (Dagan et al. (2009)) aims at capturing a broad range of inferences that are relevant for several Natural Language Processing applications, and consists of deciding, given two text fragments, whether the meaning of one text (the *hypothesis H*) is entailed, *i.e.* can be inferred, from another text (the *text T*).

Although several approaches to face this task have been experimented, and progresses in TE technologies have been shown in RTE evaluation campaigns, a renewed interest is rising in the TE community towards a deeper and better understanding of the core phenomena involved in textual inference. In line with this direction, we are convinced that crucial progress may derive from a focus on decomposing the complexity of the TE task into basic phenomena and on their combination. This belief demonstrated to be shared by the RTE community, and a number of recently published works (e.g. Sammons et al. (2010), Bentivogli et al. (2010)) agree that incremental advances in local entailment phenomena are needed to make significant progress in the main task, which is perceived as omnicomprehensive and not fully understood yet. According to this premise, the aim of this work is to systematize and delve into the work done so far in component-based TE, focusing on the aspects that contribute to highlight a common framework and to define a clear research direction that deserves further investigation.

Basing on the original definition of TE, that allows to fomulate textual inferences in an application independent way and to take advantage of available datasets for training provided in the RTE evaluation campaigns, we intend to analyse how the common intuition of decomposing TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint. Aspects related to meaning compositionality, which are absent in the original proposal, could potentially be introduced into TE and may bring new light into textual inference.

In this direction, we propose a definition for "strong" component-based TE, where each component is in itself a complete TE system, able to address a TE task on a specific phenomenon in isolation. Then, we review the literature in the TE field according to our definition, trying to position relevant work as more or less close to our idea of strong component-based TE. We have analysed and carried out research on several dimensions of the problem, including: *i)* the definition and implementation of

system components able to address specific inference types (Section 2); *ii)* the analysis of the phenomena relevant to component-based TE (Section 3); *iii)* the development of methodologies for the analysis of component-based TE systems, providing a number of qualitative indicators to assess the capabilities that systems have to address single phenomena in a pair and to combine them (Section 4).

2 Component-based TE framework

We define a component-based TE architecture as a set of clearly identifiable TE modules that can be singly used on specific entailment sub-problems and can be then combined to produce a global entailment judgement. Each component receives a certain example pair as input, and outputs an entailment judgment concerning the inference type it is built to address. In other words, each component is in turn a TE system, that performs the same task focusing only on a certain sub-aspect of entailment. According to our proposal the following requirements need to be fulfilled in component-based TE architecture: *i)* each compenent must provide a 3-way judgment (i.e. entailment, contradiction, unknown) on a specific aspect underlying entailment, where the unknown judgement might be interpreted as the absence of the phenomenon in the TE pair; *ii)* in a component-based architecure, the same inference type (e.g. temporal, spatial inferences) can not be covered by more than one component; this is because in the combination phase we do not want that the same phenomen is counted more than one time.

No specific constraints are defined with respect to how such components should be implemented, i.e. they can be either a set of classifiers or rule-based modules. In addition, linguistic processing and annotation of the input data (e.g. parsing, NER, semantic role labeling) can be required by a component according to the phenomenon it considers. An algorithm is then applied to judge the entailment relation between T and H with respect to that specific aspect. Unlike similarity algorithms, with whom algorithms performing entailment are often associated in the literature, the latter are characterized by the fact that the relation on which they are asked to judge is directional. According to such definition, the nature of the TE task is not modified, since each sub-task independently performed by the system components keeps on being an entailment task. Suitable composition mechanisms should then be applied to combine the output of each single module to obtain a global judgment for a pair.

The definition presented above provides a strong interpretation of the compositional framework for TE, that can be described as a continuum that tends towards systems developed combining identifiable and separable components addressing specific inference types. A number of works in the literature can be placed along this continuum, according to how much they get closer to this interpretation.

Systems addressing TE exploiting machine learning techniques with a variety of features, including lexical-syntactic and semantic features (e.g. Kozareva and Montoyo (2006), Zanzotto et al. (2007)) tend towards the opposite extreme of this framework, since even if linguistic features are used, they bring information about a specific aspect relevant to the inference task but they do not provide an independent judgment on it. These systems are not modular, and it is difficult to assess the contribution of a certain feature in providing the correct overall judgment for a pair. A step closer towards the direction of component-based TE is done by Bar-Haim et al. (2008), that model semantic inference as application of entailment rules specifying the generation of entailed sentences from a source sentence. Such rules capture semantic knowledge about linguistic phenomena (e.g. paraphrases, synonyms), and are applied in a transformation-based framework. Even if these rules are clearly identifiable, their application per se does not provide any judgment about an existing entailment relation between T and H.

A component-based system has been developed by Wang and Neumann (2008), based on three specialized RTE-modules: (i) to tackle temporal expressions; (ii) to deal with other types of NEs; (iii) to deal with cases with two arguments for each event. Besides these precision-oriented modules, two robust but less accurate backup strategies are considered, to deal with not yet covered cases. In the final stage, the results of all specialized and backup modules are joint together, applying a weighted voting mechanism.

Getting closer to the definition of component-based TE presented at the beginning of this Section, in Magnini and Cabrio (2009) we propose a framework for the definition and combination of specialized entailment engines, each of which able to deal with a certain aspect of language variability. A distance-

based framework is assumed, where the distance d between T and H is inversely proportional to the entailment relation in the pair. We assume an edit distance approach (Kouylekov and Magnini (2005)), where d is estimated as the sum of the costs of the edit operations (i.e. insertion, deletion, substitution), which are necessary to transform T into H. Issues underlying the combination of the specialized entailment engines are discussed, i.e. the order of application and the combination of individual results in order to produce a global result.

3 Linguistic analysis and resources for component-based TE

The idea underlying component-based TE is that each component should independently solve the entailment relation on a specific phenomenon relevant to inference, and then the judgments provided by all the modules are combined to obtain an overall judgment for a pair. Our definition abstracts from the different theories underlying the categorization of linguistic phenomena, so a straightforward relation between TE component and linguistic phenomena cannot be defined a priori. Some work has already been done in investigating in depth sub-aspects of entailment, and in developing *ad hoc* resources to assess the impact of systems components created to address specific inference types. Earlier works in the field (e.g. Vanderwende et al. (2005), Clark et al. (2007)) carried out partial analysis of the data sets in order to evaluate how many entailment examples could be accurately predicted relying only on lexical, syntactic or world knowledge. Bar-Haim et al. (2005) defined two intermediate models of textual entailment, corresponding to lexical and lexical-syntactic levels of representation, and a sample from RTE-1 data set was annotated according to each model.

A step further, other RTE groups have developed focused data sets with the aim of investigating and experimenting on specific phenomena underlying language variability. For instance, to evaluate a contradiction detection module Marneffe et al. (2008) created a corpus where contradictions arise from negation, by adding negative markers to the RTE-2 test data. Kirk (2009) describes his work of building an inference corpus for spatial inference about motion, while Akhmatova and Dras (2009) experiment current approaches on hypernymy acquisition to improve entailment classification.

The first systematic work of annotation of TE data sets is done by Garoufi (2007), that propose a scheme for manual annotation of textual entailment data sets (ARTE). The aim is to highlight a wide variety of entailment phenomena in the data, in relation to three levels, i.e. *Alignment*, *Context* and *Coreference*. 23 different features are extracted for positive entailment annotation, while for the negative pairs a more basic scheme is conceived. The ARTE scheme has been applied to the complete positive entailment RTE-2 Test Set (400 pairs), and to a random 25% portion of the negative entailment Test Set.

More recently, in Bentivogli et al. (2010) we present a methodology for the creation of specialized TE data sets, made of *monothematic T-H pairs*, i.e. pairs in which a certain phenomenon relevant to the entailment relation is highlighted and isolated (Magnini and Cabrio (2009)). Such monothematic pairs are created basing on the phenomena that are actually present in the RTE pairs, so that the distribution of the linguistic phenomena involved in the entailment relation emerges. A number of steps are carried out manually, starting from a T-H pair taken from one of the RTE data sets, and decomposing it in a number of monothematic pairs T-H_i, where T is the original text and H_i are the hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in T-H. Phenomena are grouped using both fine-grained and broader categories (e.g. *lexical*, *syntactic*, *lexical-syntactic*, *discourse* and *reasoning*). After applying the proposed methodology, all the monothematic pairs T-H_i relative to the same phenomenon *i* are grouped together, resulting in several data sets specialized for phenomenon *i*. Unlike previous work of analysis of RTE data, the result of this study is a resource that allows evaluation of TE systems on specific phenomena relevant to inference, both when isolated and when interacting with the others (the annotation of RTE data with the linguistic phenomena underlying the entailment/contradiction relations in the pairs is also provided). A pilot study has been carried out on 90 pairs from RTE-5 data set.[1]

Highlighting the need of resources for solving textual inference problems in the context of RTE, Sammons et al. (2010) challenge the NLP community to contribute to a joint, long term effort in this

[1]The resulting data sets are freely available at `http://hlt.fbk.eu/en/Technology/TE_Specialized_Data`

direction, making progress both in the analysis of relevant linguistic phenomena and their interaction, and developing resources and approaches that allow more detailed assessment of RTE systems. The authors propose a linguistically-motivated analysis of entailment data based on a step-wise procedure to resolve entailment decision, by first identifying parts of T that match parts of H, and then identifying connecting structure. Their inherent assumption is that the meanings of T and H could be represented as sets of n-ary relations, where relations could be connected to other relations (i.e. could take other relations as arguments). The authors carried out a feasibility study applying the procedure to 210 examples from RTE-5, marking for each example the entailment phenomena that are required for the inference.

4 Evaluation in component-based TE

The evaluation measure adopted in the RTE challenges is accuracy, i.e. the percentage of pairs correctly judged by a TE system. In the last RTE-5 and RTE-6 campaigns, participating groups were asked to run ablation tests, to evaluate the contribution of publicly available knowledge resources to the systems' performances. Such ablation tests consist of removing one module at a time from a system, and rerunning the system on the test set with the other modules, except the one tested. The results obtained were not satisfactory, since the impact of a certain resource on system performances is really dependent on how it is used by the system. In some cases, resources like WordNet demonstrated to be very useful, while for other systems their contribution is limited or even damaging, as observed also in Sammons et al. (2010).

To provide a more detailed evaluation of the capabilities of a TE system to address specific inference types, in Cabrio and Magnini (2010) we propose a methodology for a qualitative evaluation of TE systems, that takes advantage of the decomposition of T-H pairs into *monothematic pairs* (described in Section 3). The assumption is that the more a system is able to correctly solve the linguistic phenomena underlying the entailment relation separately, the more the system should be able to correctly judge more complex pairs, in which different phenomena are present and interact in a complex way. According to such assumption, the higher the accuracy of a system on the monothematic pairs and the compositional strategy, the better its performances on the original RTE pairs. The precision a system gains on single phenomena should be maintained over the general data set, thanks to suitable mechanisms of meaning combination. A number of quantitative and qualitative indicators about strength and weaknesses of TE systems result from the application of this methodology. Comparing the qualitative analysis obtained for two TE systems, the authors show that several systems' behaviors can be explained in terms of the correlation between the accuracy on monothematic pairs and the accuracy on the corresponding original pairs. In a component based framework, such analysis would allow a separate evaluation of TE modules, focusing on their ability to correctly address the inference types they are built to deal with.

5 Conclusions

This paper provides a definition for strong component-based TE framework, exploiting the common intuition that decomposing the complexity of TE would allow a better comprehension of the problem from both a linguistic and a computational viewpoint. We have reviewed the literature according to our definition, trying to position relevant works as more or less close to our idea of strong component-based TE. We hope that the analysis of the different dimensions of the problem we provided may bring interesting elements for future research works. In this direction, we propose a research program in which for different applications (e.g. domain, genre) specific TE component-based architectures could be optimized, i.e. composed by modules that meet the requirements of that specific genre/domain.

References

Akhmatova, E. and M. Dras (2009). Using hypernymy acquisition to tackle (part of) textual entailment. In *Proceedings of TextInfer 2009*, Singapore. 6 August.

Bar-Haim, R., J. Berant, I. Dagan, I. Greental, S. Mirkin, E. Shnarch, and I. Szpektor (2008). Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of the TAC 2008 Workshop on TE*, Gaithersburg, Maryland, USA. 17 November.

Bar-Haim, R., I. Szpektor, and O. Glickman (2005). Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan. 30 June.

Bentivogli, L., E. Cabrio, I. Dagan, D. Giampiccolo, M. L. Leggio, and B. Magnini (2010). Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of LREC 2010*, Valletta, Malta. 19-21 May.

Bentivogli, L., B. Magnini, I. Dagan, H. Dang, and D. Giampiccolo (2009). The fifth pascal recognizing textual entailment challenge. In *Proceedings of the TAC 2009 Workshop on TE*, Gaithersburg, Maryland. 17 November.

Cabrio, E. and B. Magnini (2010). Toward qualitative evaluation of textual entailment systems. In *Proceedings of COLING 2010: Posters*, Beijing, China. 23-27 August.

Clark, P., P. Harrison, J. Thompson, W. Murray, J. Hobbs, and C. Fellbaum (2007). On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-07 Workshop on TE and Paraphrasing*, Prague, Czech Republic. 28-29 June.

Dagan, I., B. Dolan, B. Magnini, and D. Roth (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE) 15*(Special Issue 04), i–xvii.

Garoufi, K. (2007). Towards a better understanding of applied textual entailment. In *Master Thesis*, Saarland University. Saarbrücken, Germany.

Kirk, R. (2009). Building an annotated textual inference corpus for motion and space. In *Proceedings of TextInfer 2009*, Singapore. 6 August.

Kouylekov, M. and B. Magnini (2005). Tree edit distance for textual entailment. In *Proceedings of RALNP-2005*, Borovets, Bulgaria. 21-23 September.

Kozareva, Z. and A. Montoyo (2006). Mlent: The machine learning entailment system of the university of alicante. In *Proc. of the second PASCAL Challenge Workshop on RTE*, Venice, Italy. 10 April.

Magnini, B. and E. Cabrio (2009). Combining specialized entailment engines. In *Proceedings of LTC'09*, Poznan, Poland. 6-8 November.

Marneffe, M. D., A. Rafferty, and C. Manning (2008). Finding contradictions in text. In *Proceedings of ACL-08*, Columbus, OH, 15-20 June.

Sammons, M., V. Vydiswaran, and D. Roth (2010). Ask not what textual entailment can do for you... In *Proceedings of ACL-10*, Uppsala, Sweden. 11-16 July.

Vanderwende, L., D. Coughlin, and B. Dolan (2005). What syntax can contribute in entailment task. In *Proceedings of the First PASCAL Challenges Workshop on RTE*, Southampton, U.K., 11-13 April.

Wang, R. and G. Neumann (2008). An accuracy-oriented divide-and-conquer strategy. In *Proceedings of the TAC 2008 Workshop on TE*, Gaithersburg, Maryland. 17 November.

Zanzotto, F., M. Pennacchiotti, and A. Moschitti (2007). Shallow semantics in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL Workshop on TE and Paraphrasing*, Prague, Czech Republic. 23-30 June.