10.12871/CLICIT2014113

An Italian Dataset of Textual Entailment Graphs for Text Exploration of Customer Interactions

Luisa Bentivogli and Bernardo Magnini
FBK, Trento, Italy
bentivo,magnini@fbk.eu

Abstract

English. This paper reports on the construction of a dataset of *textual entailment graphs* for Italian, derived from a corpus of real customer interactions. Textual entailment graphs capture relevant semantic relations among text fragments, including equivalence and entailment, and are proposed as an informative and compact representation for a variety of text exploration applications.

Italiano. *Questo lavoro riporta la costruzione di un dataset di grafi di implicazione testuale per la lingua italiana, derivati da un corpus di interazioni reali tra cliente e call centre. I grafi di implicazione testuale catturano relazioni semantiche significative tra porzioni di testi, incluse equivalenze e implicazioni, e sono proposti come un formato di rappresentazione informativo e compatto per applicazioni di esplorazione di contenuti testuali.*

1 Introduction

Given the large production and availability of textual data in several contexts, there is an increasing need for representations of such data that are able at the same time to convey the relevant information contained in the data and to allow compact and efficient text exploration. As an example, customer interaction analytics requires tools that allow for a fine-grained analysis of the customers' messages (*e.g.* complaining about a particular aspect of a particular service or product) and, at the same time, allow to speed up the search process, which commonly involves a huge amount of interactions, on different channels (*e.g.* telephone calls, emails, posts on social media), and in different languages.

A relevant proposal in this direction has been the definition of *textual entailment graphs* (Berant et al., 2010), where graph nodes represent predicates (*e.g.* $marry(x, y)$), and edges represent the entailment relations between pairs of predicates. This recent research line in Computational Linguistics capitalizes on results obtained in the last ten years in the field of *Recognizing Textual Entailment* (Dagan et al., 2009), where a successful series of shared tasks have been organized to show and evaluate the ability of systems to draw text-to-text semantic inferences.

In this paper we present a linguistic resource consisting of a collection of textual entailment graphs derived from real customer interactions in Italian social fora, which is our motivating scenario. We extend the earlier, predicate-based, variant of entailment graphs to capture entailment relations among more complex text fragments. The resource is meant to be used both for training and evaluating systems that can automatically build entailment graphs from a stream of customer interactions. Then, entailment graphs are used to browse large amount of interactions by call center managers, who can efficiently monitor the main reasons for customers' calls. We present the methodology for the creation of the dataset as well as statistics about the collected data.

This work has been carried out in the context of the EXCITEMENT project[1], in which a large European consortium aims at developing a shared software infrastructure for textual inferences, *i.e.* the EXCITEMENT Open Platform[2] (Padó et al., 2014; Magnini et al., 2014), and at experimenting new technology (*i.e.* entailment graphs) for customer interaction analytics.

[1] excitement-project.fbk.eu
[2] http://hltfbk.github.io/Excitement-Open-Platform/

2 Textual Entailment Graphs

Textual Entailment is defined as a directional relationship between two text fragments - T, the entailing text and H, the entailed text - so that *T entails H if, typically, a human reading T would infer that H is most likely true* (Dagan et al., 2006). While Recognizing Textual Entailment (RTE) datasets are typically composed of independent T-H pairs, manually annotated with "entailment" or "non entailment" judgments (see (Bentivogli et al., Forthcoming) for a survey of the various RTE datasets), the text exploration scenario we are addressing calls for a representation where entailment pairs are highly interconnected. We model such relations using *Textual Entailment Graphs*, where each node is a textual proposition (*e.g.* a predicate with arguments and modifiers), and each edge indicates a directional entailment relation.

An example of textual entailment graph is presented in Figure 1, where the node *"chi ha la chiavetta non riesce a connettersi"* entails *"non riesco a navigare con la chiavetta"*. Entailment judgments in this context are established under an existential interpretation: if there is a situation where someone *"non riesce a connettersi"*, then it is true (*i.e.* it is entailed) that, under appropriate meaning interpretation of the sentences, a situation exists in which someone *"non riesce a navigare"*. In the entailment graph, mutually entailing nodes (corresponding to pharaphrases) are represented unified in the same node, as in the case of *"chi ha la chiavetta non riesce a connettersi"*, *"la mia chiavetta non si connette"*, *"non riesco a collegarmi con la chiavetta"* in Figure 1. The graph representation also allows to derive implicit relations among nodes. For instance, since the entailment relation is transitive, the graph in Figure 1 allows to infer that *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* entails *"non riesco a navigare con la chiavetta"*. In addition, the lack of a path in the graph represents non-entailment relations, as for instance the fact that *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* does not entail *"da domenica non riesco a navigare con la chiavetta"*, because we can not establish a temporal relation between *"dal giorno 20/4"* and *"da domenica"*.

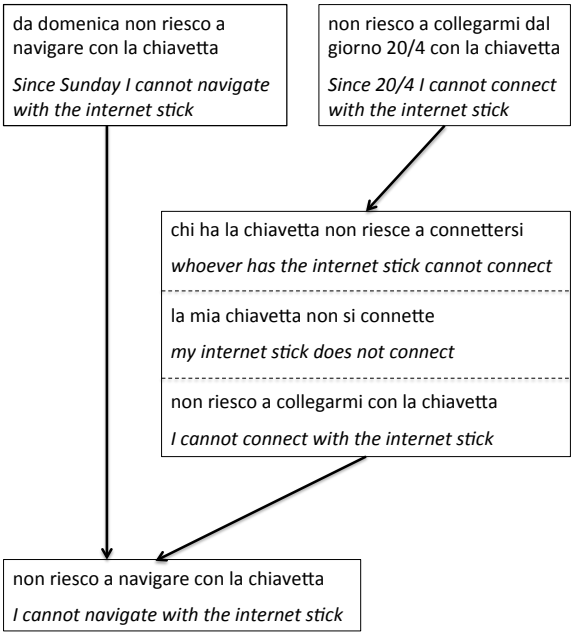

Figure 1: Portion of textual entailment graph.

3 Dataset Creation

The entailment graph creation process starts from customer interactions collected for a given topic and is composed of two main phases: (*i*) for each interaction all the relevant text fragments are extracted and the corresponding fragment graphs are created; (*ii*) all the individual fragment graphs are merged into the final entailment graph. The complete workflow of the dataset creation process is shown in Figure 2.

The starting interactions are posts taken from the official webpage of a mobile service provider in a social network, and contain reasons for dissatisfaction concerning the provider. The texts are anonymized to eliminate any reference to both the provider and the customers writing the posts.

As Figure 2 shows, the process alternates manual and automatic steps. In step 1, for each interaction the relevant text fragments are manually identified. A fragment is defined as a content unit that conveys one complete statement related to the topic (*i.e.* one reason for dissatisfaction). In our example, *"da domenica non riesco a navigare con la chiavetta"*, *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"*, *"la mia chiavetta non si connette"* are all fragments extracted from different interactions. Fragments are then generalized in order to increase the probability of recognizing entailing texts in the collection and provide a richer hierarchical structure to the entailment graph. Such generalization is performed automatically after *grammatical modifiers* of the fragments, *i.e.* tokens which can be removed

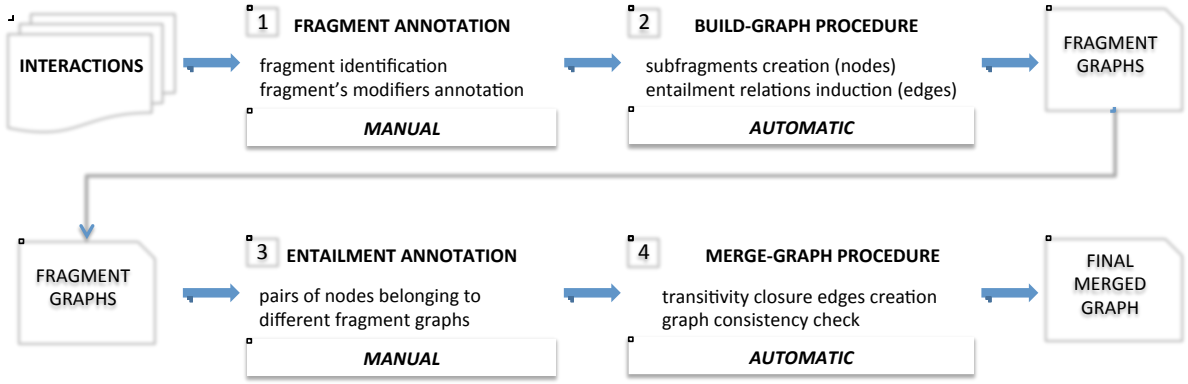

Figure 2: Entailment graph creation process.

from a fragment without affecting its comprehension, are manually specified. For example, *"da domenica"* and *"dal giorno 20/4"* are annotated as modifiers of respectively the first and second fragment above. This first manual annotation phase was carried out with the CAT Tool (Lenzi et al., 2012).[3] In step 2, given a fragment and its annotated modifiers, the corresponding subfragments are automatically created by incrementally removing its modifiers until no modifiers are left. In addition, entailment relations are automatically induced following the principle that a more specific text (*i.e.* containing more modifiers) entails a more generic one (*i.e.* containing less modifiers). As a result, an entailment graph of the corresponding fragment - Fragment Graph - is constructed, where the nodes are the fragment and its subfragments, and the edges are the entailment relations between them. In our example, for the fragment *"da domenica non riesco a navigare con la chiavetta"*, the more general subfragment *"non riesco a navigare con la chiavetta"* is automatically created as well as the entailment relation from the entailing fragment to the entailed subfragment.

To obtain the final textual entailment graph, individual fragment graphs are merged by finding all the entailment relations between their nodes. In order to minimize the number of node pairs to be manually annotated in step 3, two strategies were adopted prior to annotation, one manual and one automatic. First, clustering of fragment graphs was manually performed according to the specific topic (*i.e.* reason for dissatisfaction) expressed by the fragments. The assumption behind this strategy is that there are no entailment relations between fragment graphs belonging to different clusters (*i.e.* dealing with different reasons for dissatisfaction). As an example, two different clusters were created for fragment graphs expressing dissatisfaction about "Telefoni smartphone e cellulari" and "Consolle". The merging phase is then performed cluster by cluster, and one final merged entailment graph for each cluster is created. Second, an algorithm aimed at skipping unnecessary manual annotations is integrated in the manual annotation interface. The interface presents to annotators all the pairwise comparisons between minimal subfragments (*i.e.* texts with no modifiers). If there is no entailment relation, then all the other pairwise comparisons between the other nodes of the fragments are automatically annotated as "no entailment". If an entailment relation is annotated between minimal subfragments, then also their respective ancestors are paired and proposed for manual annotation. In our example, *"non riesco a collegarmi con la chiavetta"* is annotated as entailing *"non riesco a navigare con la chiavetta"*. Due to this entailment relation, also *"non riesco a collegarmi dal giorno 20/4 con la chiavetta"* and *"da domenica non riesco a navigare con la chiavetta"* are paired and presented for annotation, which in this case is a negative entailment judgment. Also mutual entailment can be annotated, as for *"non riesco a collegarmi con la chiavetta"*, *'chi ha la chiavetta non riesce a connettersi"*, and *"la mia chiavetta non si connette"*.

Once step 3 has been completed, in the final automatic step 4 the individual fragment graphs are merged, transitive closure edges are added, and a consistency check aimed at ensuring that there are

[3]The tool is freely available at `https://dh.fbk.eu/resources/cat-content-annotation-tool`.

Clusters	Interactions	Fragment Graphs	Total Nodes	Total Edges	Intra-Fragment Edges	Inter-Fragment Edges
19	294	344	760	2316	733	1583

Table 1: Composition of the dataset.

no transitivity violations is carried out.

As a result of fragment graph merging, a textual entailment graph over the input fragments is constructed.

Statistics about the composition of the dataset created according to the described procedure are presented in Table 1. The final dataset contains 19 consistent textual entailment graphs, one for each of the clusters into which the fragment graphs were subdivided. The table also shows the number of original interactions and the fragment graphs derived from them (step 1 of the process), and the total number of nodes and edges composing the 19 final entailment graphs resulting from the merging of fragment graphs (step 4 of the process). Finally, the total number of edges contained in the final graphs is further subdivided into intra-fragment and inter-fragment edges. Intra-fragment edges denote edges connecting the nodes within fragment graphs, *i.e.* edges generated during fragment graph construction. Inter-fragment edges are edges generated during the merge phase.

The dataset is released for research purposes under a Creative Commons Attribution-NonCommercial-ShareAlike license, and will be available at the EXCITEMENT project website by the end of the project (31/12/2014). The release will also contain information about Inter-Annotator Agreement, which is being currently calculated for the two manual annotation phases carried out during dataset creation, namely (*i*) the identification of modifiers within text fragments, which is necessary to build the fragment graphs (step 1 of the process), and (*ii*) the annotation of entailment relations between statements (nodes) belonging to different fragment graphs, which is required to merge the fragment graphs (step 3).

4 Conclusion

We have presented a new linguistic resource for Italian, based on textual entailment graphs derived from real customer interactions. We see a twofold role of this resource: *(i)* on one side it provides empirical evidences of the important role of semantic relations and provides insights for new developments of the textual entailment framework; *(ii)* on the other side, a corpus of textual entailment graphs is crucial for the realization and evaluation of automatic systems that can build entailment graphs for concrete application scenarios.

Acknowledgments

This work was partially supported by the EC-funded project EXCITEMENT (FP7ICT-287923).

References

Luisa Bentivogli, Ido Dagan, and Bernardo Magnini. Forthcoming. The recognizing textual entailment challenges datasets. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global learning of focused entailment graphs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1220–1229, Uppsala, Sweden.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering*, 15(4):i–xvii.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of the 52nd Meeting of the Association for Computational Linguistics, Demo papers*.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*. doi:10.1017/S1351324913000351.