

Content Extraction Meets the Social Web in the LiveMemories Project

Massimo Poesio
University of Trento

Bernardo Magnini
FBK-irst, Trento, Italy

Introduction

The widespread availability of low-cost means for putting into digital form multimodal information including text of both traditional and non-traditional type (from stories to blogs), photos, and videos, together with the explosion in use of social networking sites such as Facebook for publishing such information about oneself or about the events of the day on the Web and sharing it with friends and perfect strangers, is leading to radically new forms for the preservation of information and the creation of collective memory.

However, the functionality offered by current social networking sites does not go beyond the upload and indexing of such information; indexing is most often word-based or at most topic-based, and the data thus collected lie otherwise unanalyzed. The aim of LiveMemories¹ is to take advantage of techniques for extracting content from multimedia sources to make such shared digital repositories ‘alive’ by identifying people and objects mentioned in them, and extracting information about events and other relations between objects, including temporal information.

This type of analysis will enable new presentation methods. Consider the following scenario. Luisa Tomasi from Gardolo goes to a Franco Battiato concert in Trento on February 17th and 18th, and takes some pictures. The next day she creates a description of the event on the LiveMemories portal, uploading the images she took and accompanying them with text describing her experience and giving her comments on the concert. Images and text are analyzed by the LiveMemories platform, that recognizes the event as one listed on www.crushsite.it and identifies Franco Battiato as one of the individuals stored in its knowledge base (automatically extracted by processing Wikipedia text and text from the local press) indeed. LiveMemories can then offer to Luisa further information about the event, e.g., it can tell Luisa that the brilliant viola player is called Demetrio Comuzzi, or it may offer to Luisa to visualize Battiato’s discography in the form of a chronology to discover when a particular song came out. LiveMemories may also discover that other people with an account on the portal went to that same concert including e.g. Mario Boato, who also uploaded his

¹LiveMemories is a three years project funded by the Autonomous Province of Trento. The project, started in October 2008, is a collaboration among three academic partners, Fondazione Bruno Kessler (FBK), University of Trento (Italy), and University of Southampton (UK), and a number of companies and data providers located in the Trentino area. Detailed information can be found at the project web site: <http://www.livememories.org>.

own data. LiveMemories can point this out to Luisa and Mario, who may also discover they share a preference for Battiato's early music.

LiveMemories includes activities in Content Extraction from text and images, Content Presentation, and Content Integration. In this paper, we will focus on Content Extraction from text and on cross-document coreference.

Background

The availability of high-performance tools for POS tagging and parsing has made it possible to contemplate large-scale semantic processing (named entity extraction, coreference, relation extraction, ontology population). US initiatives such as MUC, ACE and GALE made large annotated resources available and introduced quantitative evaluation.

In intra-document coreference (IDC) this led to the development of the first large-scale machine learning models using these resources and to the development of IDC tools, most recently, the ELKFED/BART system (Versley et al., 2008). In relation extraction, work carried out as part of the ACE initiative and in ELERFED showed that good results can be obtained extracting relations from news with supervised methods, particularly Support Vector Machines (SVMs) and Kernel Methods but that semi-supervised methods are more effective with less formal text.

Interest in cross-document coreference (CDC) has began fairly recently (Bagga and Baldwin, 1998), but there has been much development in recent years because of great interest both from government and from industry. In particular there has been great interest in a simpler form of entity disambiguation, generally known as Web entity as in the case of the Web people task of Semeval (Popescu and Magnini, 2007) and the Spock challenge². As testified by the SEMEVAL Web People task³, most state of the art systems are based on clustering of entity descriptions containing a mixture of collocational and other information, among which information about entities and relations. SEMEVAL also showed that the clustering technique and especially the termination criterion are crucial. Finally, work on the Spock challenge highlighted the need for methods for handling huge quantities of information. Recent developments have therefore focused on improving the clustering technique and experimenting with different types of information that can be extracted robustly from text. (See, e.g., the results with ELERFED.) Most of the work discussed above was carried out for English; progress with languages other than English includes work on German (e.g., Versley) and Spanish (Ferrandez) but very little on Italian apart from work by Delmonte, also in part for lack of resources. In this direction it is worth to mention the creation of a reference benchmark for CDC in Italian (see Bentivogli et al. (2008)), which is used and improved in the LiveMemories project.

²challenge.spock.com/.

³<http://nlp.uned.es/weps/>.

Goals of LiveMemories: Content Extraction from Text

LiveMemories builds on top of previous content extraction technologies. Particularly, we use TextPro⁴ (Pianta et al. (2008)) a suite of modular Natural Language Processing tools for analysis of Italian and English texts. The current version of the tool suite provides functions ranging from tokenization to chunking and Named Entity Recognition. TextPro performed the best on the task of Italian NER and Italian PoS Tagging at EVALITA 2007⁵.

The performance and usefulness of existing technology for content extraction from text is currently improved by:

- Larger corpora and techniques, obviating the need for large scale annotation (e.g., active learning, weakly supervised methods).
- Better preprocessing techniques (often underestimated);
- Incorporating automatically extracted lexical and commonsense features in addition to traditional 'surface' features;
- Developing better Machine Learning methods to exploit these more advanced sources of information (e.g. kernel functions)
- Developing richer representations of relations, e.g., with temporal modification (e.g. *John Doe was CFO of ACME from 2001 to 2005*);
- Further developing automatic methods for Textual Entailment Recognition, a robust type of textual inference based on patterns that can be automatically acquired from corpora. We use the EDITS system⁶ (Negri et al. (2009)).

Conclusions

In the first year of the project, we made substantial progress in content extraction, including:

- The development of a new cross-document coreference resolver based on the work by Popescu (Popescu and Magnini, 2007);
- The creation of a new annotated corpus for coreference in Italian;
- The development of a new intra-document coreference resolver for Italian based on BART;

In addition, we made lot of progress on building a user community, establishing contact with a number of communities in Trentino that will become pilot users of our platform; and designed a new platform for digital memories centered around the notion of **story**. The first release of the platform will become available end of September 2009.

⁴TextPro is freely distributed for research purposes at <http://textpro.fbk.eu/>.

⁵<http://evalita.fbk.eu/>.

⁶EDITS is distributed as open source software at <http://edits.fbk.eu/>.

References

- Bagga, A. and Baldwin, B.(1988), *Entity-Based Cross-Document Coreferencing Using the Vector Space Model*, in *Proceedings of COLING/ACL*, Montreal, 1998.
- Bentivogli L., Girardi C. and Pianta E. (2008), *Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News*, in *Proceedings of LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, Marrakech, Morocco.
- Negri M., Kouylekov M., Magnini B., Mehdad Y. and Cabrio E. (2009) *Towards Extensible Textual Entailment Engines: the EDITS Package*, in *Proceedings of the XI Conference of the Italian Association for Artificial Intelligence - to appear*, Reggio Emilia, Italy.
- Pianta E., Girardi C. and Zanolini R. (2008), *The TextPro tool suite*, in *Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco.
- Popescu, O. and Magnini, B. (2007) *IRST-BP: Web People Search Using Name Entities*, in *Proceedings of SEMEVAL*, 2007.
- Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X. and Moschitti, A. (2008), *BART: A modular toolkit for coreference resolution*, in *Proceedings of LREC*, Marrakesh, 2008.