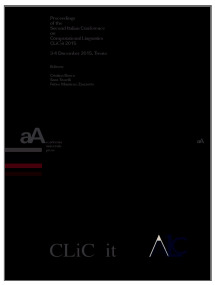

Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC)

Manuela Speranza and Anne-Lyse Minard

DOI: 10.4000/books.aaccademia.1550
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: November 11, 2016
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788899200008

Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC)

Manuela Speranza, Anne-Lyse Minard
Fondazione Bruno Kessler, Trento
{manspera,minard}@fbk.eu

Abstract

English. In this paper we present the annotation of events, entities, relations and coreference chains performed on Italian translations of English annotated texts. As manual annotation is a very expensive and time-consuming task, we devised a cross-lingual projection procedure based on the manual alignment of annotated elements.

Italiano. *In questo articolo descriviamo l'annotazione degli eventi, delle entità, delle relazioni e delle catene di coreferenza realizzata su traduzioni in italiano di testi inglesi già annotati. Essendo l'annotazione manuale un compito molto dispendioso, abbiamo ideato una procedura di proiezione interlinguale basata sull'allineamento degli elementi annotati.*

1 Introduction

The NewsReader Wikinews Italian Corpus (WItaC) is a new Italian annotated corpus consisting of English articles taken from Wikinews[1] and translated into Italian by professional translators.

The English corpus was created and annotated manually within the NewsReader project,[2] whose goal is to build a multilingual system able to reconstruct storylines across news articles in order to provide policy and decision makers with an overview of what happened, to whom, when, and where. Semantic annotations in the NewsReader English Wikinews corpus span over multiple levels, including both intra-document annotation (entities, events, temporal information, semantic roles, and event and entity coreference) and cross-document annotation (event and entity coreference). As manual annotation is a very expensive and time-consuming task, we devised a procedure to automatically project the annotations already available in the English texts onto the Italian translations, based on the manual alignment of the annotated elements in the two languages.

The English corpus, taken directly from Wikinews, together with WItaC, being its translation, ensures access to non-copyrighted articles for the evaluation of the NewsReader system and the possibility of comparing results in the two languages at a finegrained level.

WItaC aims at being a reference for the evaluation of storylines reconstruction, a task requiring several subtasks, e.g. semantic role labeling (SRL) and event coreference. In addition, it is part of a cross-lingually annotated corpus,[3] thus enabling for experiments across different languages.

The remainder of this article is organized as follows. We review related work in Section 2. In Section 3 we present the annotations available in the English corpus used as the source for the projection of the annotation. In Section 4 we detail some adaptations of the guidelines specific for Italian. In Sections 5 and 6 we describe the annotation process and the resulting WItaC corpus. Finally, we conclude presenting some future work.

2 Related work

A number of semantically annotated corpora are available for English, whereas most other languages are under-resourced. As far as Italian is concerned, WItaC is the first corpus offering annotations of entities, events, and event factuality, together with semantic role labeling and cross-document coreference annotation.

For entities and entity coreference, the reference Italian corpus is I-CAB (Magnini et al., 2006),

[1] Wikinews (http://en.wikinews.org) is a collection of multilingual online news articles written collaboratively in a wiki-like manner.

[2] http://www.newsreader-project.eu/

[3] The NewsReader consortium has annotated also the Spanish and Dutch translations of the same Wikinews articles.

which is annotated with entities, time expressions (following the TIMEX2 standard), and intra-document entity coreference; for cross-document person entity coreference, we refer to CRIPCO (Bentivogli et al., 2008). Regarding temporal information and event factuality, two annotated corpora are available: respectively, the EVENTI corpus (Caselli et al., 2014), used as the evaluation dataset for the EVENTI task at Evalita 2014 and annotated with events, time expressions (TIMEX3), temporal signals, and temporal relations, and FactIta Bank (Minard et al., 2014), a subsection of EVENTI annotated with event factuality.

To the best of our knowledge there exist no other Italian corpora with semantic role labeling and event cross-document coreference annotation. The reference corpus for SRL in English is the CoNLL-2008 corpus (Surdeanu et al., 2008). For cross-document coreference, the ECB+ corpus (Cybulska and Vossen, 2014) has recently been created extending the ECB corpus.

The method we propose for cross-lingual annotation projection taking advantage of the alignment between texts in two different languages is similar to other methods used, for example, to build annotated corpora with semantic roles (Padó and Lapata, 2009), temporal information (Spreyer and Frank, 2008; Forascu and Tufi, 2012), and coreference chains (Postolache et al., 2006). However, previous work is based on the use of corpora aligned at the word level either manually, which is very time-consuming, or automatically, which is error prone. On the other hand, our method envisages a manual alignment at the markable level, where the extent of each element is annotated on the translated text and then aligned to the English annotated element on a semantic rather than syntactic basis.

3 Annotation available in the English source corpus

The NewsReader Wikinews English corpus contains intra-document semantic annotation and cross-document coreference annotation.

3.1 Annotation at document level

The annotation is based on the NewsReader guidelines (Tonelli et al., 2014) and was performed using the CAT tool (Bartalesi Lenzi et al., 2012). The first five sentences (including the headline) of each document contain the following annotations: markables, relations, and intra-document coreference.

Markable annotation. Textual realizations of entity instances, referred to as entity mentions, are the portions of text in which entity instances of different types (people, organizations, locations, financial entities, and products) are referenced within a text. Each entity mention is described through that portion of text (extent) and two optional attributes, i.e. syntactic head and syntactic type.

The textual realization of an event, the event mention, can be a verb, a noun, a pronoun, an adjective, or a prepositional construction. It is annotated through its extent and a number of attributes, e.g. predicate (lemma), part-of-speech, and factuality. Factuality attributes (van Son et al., 2014) of an event include time, certainty and polarity.

The annotation of temporal expressions is based on the ISO-TimeML guidelines (ISO, 2012) , and thus includes durations, dates (e.g. the document creation time), times, and sets of times, with the following attributes : type, normalized value, anchorTimeID (for anchored temporal expressions), and beginPoint and endPoint (for durations).

Numerical expressions include percentages, amounts described in terms of currencies, and general amounts. Temporal signals, inherited from ISO-TimeML, make explicit a temporal relation. Similarly, causal signals (C-SIGNALS) indicate the presence of a causal relation between two events (e.g. *because of*, *since*, *as a result*, and *the reason why*).

Relation annotation. Based on the TimeML approach (Pustejovsky et al., 2003), temporal relations (e.g. 'before', 'after', 'includes', and 'ends') are used to link two event mentions, two temporal expressions or an event mention and a temporal expression. The annotation of subordinating relations also leans on TimeML, although its scope was reduced to the annotation of reported speech.

In addition, explicit causal relations between causes and effects denoted by event mentions have been annotated taking into consideration the *cause*, *enable*, and *prevent* categories of causation, and grammatical relations have been created for events that are semantically dependent on another event, to link them to their governing content verb/noun.

Semantic role labeling is modeled through the HAS_PARTICIPANT relation, a one-to-one relation linking an event mention to an entity mention playing a role in the event. PropBank (Bonial et al., 2010) is used as the reference framework for the assignment of the semantic role to each relation.

Intra-document event and entity coreference. The annotation of coreference chains that link different mentions to the same instance is based on the REFERS_TO relation.

Entity instances are described through the non text-consuming ENTITY tag and the two attributes entity type and tag descriptor; similarly, event instances are described through the non text-consuming EVENT tag and the two attributes event class and tag descriptor.

3.2 Annotation at corpus level

Annotation at the corpus level (Speranza and Minard, 2014), performed using the CROMER tool (Girardi et al., 2014), relies on the creation of corpus instances (both entities and events) and on links holding between each mention and the corpus instance it refers to. Corpus instances are described through a unique instance ID and the DBpedia URI (when available). Annotation consists of:

- cross-document entity coreference in the first five sentences;
- cross-document entity and event coreference in the whole document for a subset of 44 seed entities (i.e., annotation and coreference of all mentions referring to the seed entities and of the events of which the entities are participants).

4 Italian language specific annotations

We adopted the NewsReader guidelines already available for English with some minor language specific adaptations, as described in detail in Speranza et al. (2014). For this reason the data on inter-annotator agreement provided for English by van Erp et al. (2015) can be used as a reference.

For the annotation of clitics, which do not exist in English, we decided to leave the annotation at the word level, rather than split it into smaller units, so as to be consistent with annotations on existing corpora, e.g. I-CAB (Magnini et al., 2006). So in the case of a token composed of a verb (i.e. an event mention) and a clitic corresponding to a pronominal mention of a markable entity, the whole token was annotated both as an entity and as an event. The syntactic head attribute of the entity mention, having as value the clitic, and the predicate attribute of the event mention, having as value the verbal root, contribute to distinguish the two annotated elements (see [1]).

(1) *Aveva già deciso di dargli un aiuto* ('He had already decided to give him some help')

EVENT_MENTION: [dargli], `pred` "dare"
ENTITY_MENTION: [dargli], `head` "gli"

As Italian, unlike English, is a null-subject language where clauses lacking an explicit subject are permitted, we devised specific guidelines that allowed us to straightforwardly align English pronouns to Italian null subjects. In particular, null subjects having finite verb forms as predicates and referring to existing entity instances (see [2]) were marked through the creation of an empty (i.e. non text-consuming) ENTITY_MENTION tag, which was then linked to other markables following the guidelines for regular text consuming entity mentions; in addition, annotators filled the tag descriptor attribute with a human friendly name and the sentence number (e.g. "He-LuiS2" for the null subject in [2]).

(2) *Obama fece un discorso. [Ø] Disse che [...]* ('Obama gave a speech. [He] said that [...]')

The annotation of modals for Italian is based on It-TimeML, where they are marked as events like all other verbs.[4]

5 Annotation of WItaC

The method we propose for the annotation of the Italian corpus consists of cross-lingual projection of annotation from a source corpus to a target corpus; it enabled us to reduce the effort by approximately three times. The annotation was performed in five steps starting with a file containing the source English annotated text and the Italian translation aligned at the sentence level.

1. Mention annotation. The first step of the annotation, performed using the CAT tool, consisted of the identification and annotation of all markable extents.

2. Alignment. The use of CAT, which is highly customizable, enabled us to set up the alignment between Italian and English markables by simply adding to the Italian markables a new attribute which takes as value the ID of a different markable. Annotators filled this attribute with the corresponding English markable by using drag-and-drop. In some cases it was also necessary to mark the attributes and/or relations that should not be imported (by writing a note in the comment attribute), or to create extra relations.[5] If a mention had no equivalent, annotators filled in the values of the attributes

[4] In WItaC modals are also linked to their governing verb through a grammatical link.

[5] No exceptions were needed for aligning null subjects.

and created the relations in which it was involved and, if it did not already exist, the instance to which it referred.

3. Automatic projection. The automatic projection was performed using a Python script working on the XML files produced by the CAT annotation tool. For each article, the script takes as input the file containing both the English fully annotated text and the Italian text on which the annotated markables have been aligned. It produces as output a file in which the Italian text has been enriched with the annotations imported from English, i.e. the event instances, the entity instances, the relations (including the REFERS_TO relation which models intra-document coreference), and the values of the non-language-specific attributes (unless a specific comment is present).

4. Manual revision. Manual revision consists of an overall check of the annotations imported automatically; in particular, it involves the annotation of the language specific attributes and the deletion of the relations that had been marked as non-importable (using the CAT tool).

5. Projection of cross document coreference. The projection of the cross-document annotation consists of importing coreference from the English corpus taking advantage of the alignment performed in the second step and extending the entity and event instances by importing the IDs of the English instances and their DBpedia URIs.

6 Dataset Description

WItaC is composed of 120 articles. In Table 1 we give the size of the whole corpus and the size of the "first 5 sentences" section, i.e. the subsection annotated with markables, relations, intra-document coreference and cross-document entity coreference. In total 6,127 markables have been annotated in Italian; of these, 5,580 are aligned to English markables while 547 have no English correspondent.

	Whole corpus		First 5 sentences	
	Ita.	Eng.	Ita.	Eng.
# files	120	120	120	120
# sentences	1,845	1,797	597	597
# tokens	44,540	40,231	15,676	13,981

Table 1: Italian and English corpus size

Exploiting the alignment, relations and attributes have been imported automatically. For only 5.7% of the markables the attributes could not be projected (e.g. two events with different PoS). In Table 2 we present the number of markables and relations annotated in the Italian corpus. Out of the total 2,709 entity mentions, 56 are null subjects aligned with English pronominal entity mentions.

Markables		Relations	
EVENT_MENTION	2,208	SLINK	220
ENTITY_MENTION	2,709	TLINK	1,711
TIMEX3	507	CLINK	61
VALUE	415	GLINK	300
SIGNAL	253	HAS_PART	1,865
C-SIGNAL	35		
Total	6,127	Total	4,157
Instances		**Coreference chains**	
EVENT INSTANCE	1,773	REFERS_TO	3,054
ENTITY INSTANCE	1,281		
Total	3,054		

Table 2: Annotations in the first five sentences

As a result of the projection of event and entity cross-document coreference chains from English, WItaC contains 740 entity instances and 887 event instances annotated at the corpus level. Annotation by projection enables us to also have cross-lingual annotation, which means that the instances are shared between English and Italian.

7 Conclusions and future work

We have presented WItaC, a new corpus consisting of Italian translations of English texts annotated using a cross-lingual projection method. We acknowledge some influence of English in the translated texts (for instance, we noticed an above-average occurrence of noun modifiers, as in "dipendenti Airbus") and in the annotation (for instance, annotators might have been influenced by English in the identification of light verb constructions in the Italian corpus). On the other hand, this method enabled us not only to considerably reduce the annotation effort, but also to add a new cross-lingual level to the NewsReader corpus; in fact, we now have two annotated corpora, in English and Italian, in which entity and event instances (in total, over 1,600) are shared.

In the short-term we plan to manually revise the projected relations and add the language-specific attributes. We also plan to use the corpus as a dataset for a shared evaluation task and afterwards we will make it freely available from the website of the HLT-NLP group at FBK[6] and from the website of the NewsReader project.

[6] http://hlt-nlp.fbk.eu/technologies.

Acknowledgments

This research was partially funded the EU NewsReader project (FP7-ICT-2011-8 grant 316404).

References

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. 2008. Creating a gold standard for person cross-document coreference resolution in italian news. In *LREC Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. PropBank Annotation Guidelines, Version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita. In *Proceedings of the First Italian Conference on Computational Linguistic CLiC-it 2014 & the Fourth International Workshop EVALITA 2014 Vol. II: Fourth International Workshop EVALITA 2014*, pages 27–34.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Corina Forascu and Dan Tufi. 2012. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. 2014. CROMER: A Tool for Cross-Document Event and Entity Coreference. In *Proceedings of the 9th Internation Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

ISO. 2012. *ISO 24617-1: Language Resource Management. Semantic Annotation Framework (SemAF). Time and Events (SemAF-Time, ISO-TimeML)*. ISO International Standard.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of the First Italian Conference on Computational Linguistic CLiC-it 2014*, pages 260–264.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research*, 36(1):307–340, September.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34.

Manuela Speranza and Anne-Lyse Minard. 2014. NewsReader Guidelines for Cross-Document Annotation. Technical Report NWR2014-9, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-9.pdf`.

Manuela Speranza, Ruben Urizar, and Anne-Lyse Minard. 2014. NewsReader Italian and Spanish specific Guidelines for Annotation at Document Level. Technical Report NWR2014-6, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/02/NWR-2014-61.pdf`.

Kathrin Spreyer and Anette Frank. 2008. Projection-based Acquisition of a Temporal Labeller. In *Proceedings of IJCNLP*, pages 489–496, Hyderabad, India, January.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings*

of the Twelfth Conference on Computational Natural Language Learning, CoNLL '08, pages 159–177, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf.

Marieke van Erp, Piek Vossen, Rodrigo Agerri, Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Annotated Data, version 2. Technical Report D3-3-2, VU Amsterdam. http://www.newsreader-project.eu/files/2012/12/NWR-D3-3-2.pdf.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).