Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus

Kepa J. Rodríguez [a], Francesca Delogu [a], Yannick Versley [b], Egon W. Stemle [a], Massimo Poesio [a]

(a) University of Trento, Italy. (b) University of Tübingen, Germany.

Introduction

The Live Memories corpus is an Italian corpus annotated with anaphoric relations. This annotation effort aims to contribute to two significant issues for CL research: the **lack of annotated anaphoric ressources for Italian** and the **increasing interest for social media data**.
The target of our annotation effort is to build a corpus of 450,000 words in 3 datasets: Wikipedia sites about Trentino / Südtirol, blog sites with users' comments and articles of the local news paper l'Adige.

The corpus includes human annotation of **morphosyntactic agreement**, **semantic class**, **anaphoricity** and **ambiguity**. The annotation scheme takes into account specific phenomena of the Italian, like the **empty subjects** and the **attached clitics**.
We plan to distribute the Wikipedia and blogs datasets under the Creative Commons Attributions (CC) license.

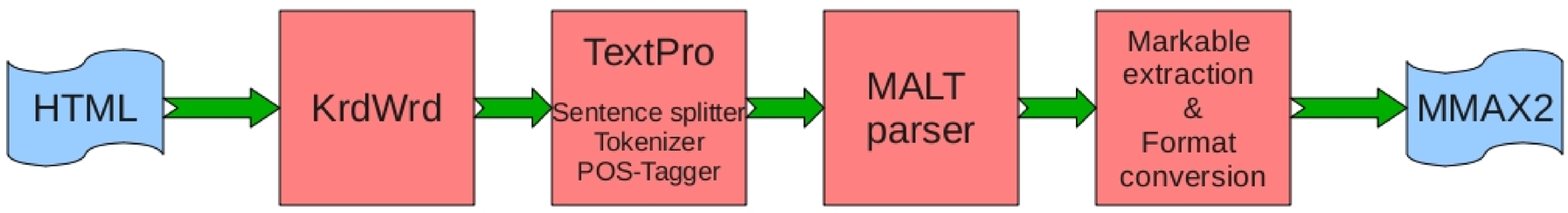
Annotation Methodology

Selection of the texts
We selected texts from pages created under the CC license, that are related to the region Trentino/Südtirol, like villages, relevant people, artworks, local facilities, etc. Selected texts have a length between 400 and 2500 words. Often Wikipedia texts are too long for the human annotation. In this case we select for the annotation parts of the text taking into account that no pronouns refer to entities outside of the selection.

Extraction of the texts
We annotate the web pages using the KrdWrd tool. This tool allows to extract the text for further processing and to keep information about the layout of the web page.

Extraction of the markables
We pre-process the text with tokenizer, sentence spliter and POS tagger from the TextPro toolkit. Then we parse the data using MALT dependency parser and use the parse trees to create markables for all NPs of the text. Finally, we convert the output of the preprocessing pipeline into the XML format used by MMAX2.

Human annotation
The first step consists of the manual correction of possible errors in the markable boundaries and of the addition of non-recognized markables. The main corrections are the identification and annotation of clitics and empty subjects, and the identification of discontinuous markables. After the markable boundaries have been corrected each markable is annotated with the tags provided in the annotation scheme.



Annotation Scheme

All NPs are treated as markables for the annotation and are annotated with morphosyntactic (`gender`, `number`, and `person`) and semantic attributes. The semantic annotation starts with a **reference** attribute that can take the values `non-referring`, `discourse-new`, or `discourse-old`.

`non-referring` markables can take the values `expletive`, `idiom`, `quantifier`, `coordination`, and `predicate`.

Coordination and discontinuous markables
Discontinuous markables are used in cases of coordinations of modified NPs where the semantic unit of one NP is interrupted, as in 1

(1) **`[Enrico]`** and [Elsa **`[Conci]`**].

Semantic category
`discourse-new` and **`discourse-old`** markables are annotated with the tagset: `Person`, `Organization`, `GPE`, `Location`, `Facility`, `Temporal`, `Numerical`, `Animate`, `Concrete`, and `Abstract`.

Anaphoric links
`discourse-old` markables are linked to their most recent antecedent. Information about the anaphoric relation is annotated with:

- Type of reference, to identify cases of discourse-deixis. No link is realized between the referring NP and the antecedent clause.
- Phrase antecedent: simple or multiple antecedent
- Ambiguity, to identify cases of referential ambiguity between two or more interpretations.

Bridging relations
`discourse-old` and `discourse-new` markables are also associated with a `related-object` attribute that indicates whether the NP stands in a bridging relation (`part-of`, `set-membership`, `attribute`) with a previously introduced NP.

Annotation of empty subjects and incorporated clitics
A `markable_type` attribute is introduced to specify whether a markable contains a nominal phrase or a verbal form standing for an empty subject or an incorporated clitic. Verbal markables are associated with a Verbal type attribute to specify whether the markable is a clitic or an empty subject. All verbal markables are then annotated with the same set of attributes used for nominal markables.

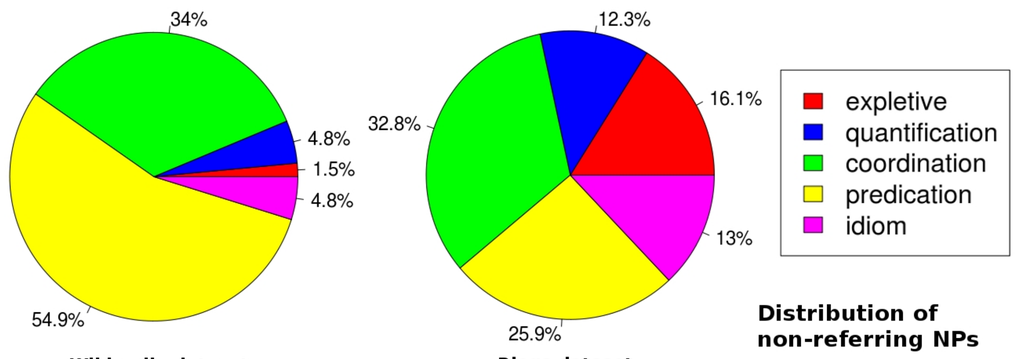
(2) [Cesare Battisti]... è stato un [geografo, politico e irredentista italiano]. [Nacque] in [Trentino]...
Cesare Battisti... was an Italian gegrapher, politician and 'irredentista'. (He) was born in Trentino...

Description of the Corpus

Wikipedia dataset
Currently annotated 144 files with 142,000 words and 44,500 markables. 57.8% of the markables are discourse-new, 28.5% discourse-old, and 13.7% are non-referring expresions mostly, cases of predication.

Blogs dataset
Currently annotated 66 files with 50,000 words and 14,000 markables. 64.7% of the markables are discourse-new, 23.6% discourse-old and 11.7% non-referring-expressions mostly, cases of coordination and predication.

Main differences
The main difference between both datasets is the distribution of the categories of non-referring NPs.



Inter-annotator agreement

Annotation on markable level
Basic annotation of the markable: `discourse-new`, `segment-antecedent`, `phrase-antecedent`, `expletive`, `quantifier`, `predicate`, `coordination` and `idiom`. κ = 0.79
Semantic type: κ = 0.85

Annotation of anaphoric links
Link to antecedent: κ = 0.88
Antecedent of clitics: κ = 0.84
Antecedent of empty subject: κ = 0.93

Acknowledgements

This research was funded by Autonomous Province of Trento - PAT, through the Grande Progetto LiveMemories.