Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus

Kepa J. Rodríguez♣ , Francesca Delogu♣,Yannick Versley♠, Egon W. Stemle♣, Massimo Poesio♣

♣University of Trento
♠University of Tübingen
♣{kepa.rodriguez | francesca.delogu | egon.stemle | massimo.poesio}@unitn.it,
♠versley@sfs.uni-tuebingen.de

Abstract

The Live Memories corpus is an Italian corpus annotated for anaphoric relations. The corpus includes manual annotated information about morphosyntactic agreement, anaphoricity, and semantic class of the NPs. For the annotation of the anaphoric links the corpus takes into account specific phenomena of the Italian language like incorporated clitics and phonetically non realized pronouns. The Live Memories Corpus contains texts from the Italian Wikipedia about the region Trentino/Süd Tirol and from blog sites with users' comments. It is planned to add a set of articles of local news papers.

1. Introduction

The social Web has become increasingly interesting for CL research. Socially constructed encyclopedias such as Wikipedia are at the moment perhaps the most important source of knowledge for NLP applications ((Ponzetto and Strube, 2007); (Mihalcea, 2007)).

Blogs are a key source of opinions and subjective judgments on a variety of topics. At this point, however, the resources needed to develop methods for handling social media data are missing: to our knowledge, there are no annotated corpora that can be used to train methods for entity extraction and relation extraction on these types of text (although there are of course resources for training opinion miners (Bo and Lee, 2008)). In the work discussed here we aim at fixing this problem. Furthermore, we provide an annotated corpus for Italian, a language with scarce language resources.

The LiveMemories corpus (henceforth LMC) is a new annotated corpus of Italian under construction as part of work on the LiveMemories project (Magnini and Poesio, 2009). The corpus will eventually include texts from Wikipedia, blogs, and news articles. This article presents the current state of ongoing work on anaphoric annotation.

2. Related work

2.1. Analyzing social web text

Most of the work done using data from blogs and other social media and networks focuses on opinion and sentiment mining.

(Hopkins and King, 2010) presents and compares new methods of automated content analysis able to be trained with a small amount of hand coded data and to classify potential large sets of documents. The presented methods have been evaluated doing analysis and measure of political opinions posted on blogs.

(Boldrini et al., 2009) presents *EmotiBlog*, a multilingual annotation scheme for social web texts like blogs and forums. *EmotiBlog* has been used to produce a dataset of English, Italian, and Spanish annotated data. The scheme classifies the sentences in objective and subjective speech. Objective speech is annotated with the source (or speaker), and subjective speech with associated emotion (like criticism, happyness, surprise, etc.) and polarity (positive or negative emotion). Then a more fine grained annotation is done at the word / collocation level. The scheme gives instructions to mark anaphoric items with source and type of anaphora. This dataset has been used to study the combination of factual and opinion question answering (Balahur et al., 2009a).

A subset of *EmotiBlog* has been used to annotate a collection of 51 blog posts with users' comments about five different topics (economy, science and technology, cooking, society and sport). Sentences that express a subjective opinion are annotated with polarity, function that indicates whether the opinion expressed is positive or negative, level or intensity of the polarity, source of the text and target or topic. This annotation was used to train and evaluate an opinion summarizer (Balahur et al., 2009b).

(Nicolov et al., 2008) investigates the **use of coreference information** for opinion mining of blog data. The algorithm uses anaphoric information of nominals and pronouns. The results show that the use of coreference information increases the performance of the system in a 10%.

(Hendrickx and Hoste, 2009) investigates the performance of an automatic coreference resolver trained with the KNACK-2002 Dutch corpus of news paper articles. The resolver was evaluated on a set of news paper texts and on a set of commented blog data. The performance decreases dramatically when the system tries to resolve coreference in non conventional text like users' comments.

The results and conclusions reported in both (Nicolov et al., 2008) and (Hendrickx and Hoste, 2009) suggest the necessity of specific sets of blog data annotated with anaphoric

information.

2.2. Anaphoric annotation

The anaphoric annotation in the LMC builds, on the one hand, on the methods for anaphoric annotation developed for the ARRAU corpus of anaphoric information in English (Poesio and Artstein, 2008) and, on the other hand, on the methods for the anaphoric annotation of Italian developed in the VENEX project (Poesio et al., 2004)[1].

The ARRAU corpus, like the LMC, consists of texts from different genres, including dialogues from the Trains Corpus, news articles from the Wall Street Journal set of the Penn Treebank, narrative texts, and art history texts from the Gnome Corpus. ARRAU studied ambiguity in anaphoric expressions and the ARRAU coding scheme allows for the explicit representation of multiple antecedents when the annotator cannot find a clearly identifiable unique antecedent. The scheme also includes methods for marking reference to abstract objects. However, the key aspects of the ARRAU coding scheme we utilize are: the methods for marking referring and non referring nominals, syntactic and semantic agreement features, and bridging.

VENEX (the Venice Essex Corpus), a joint project between the Universities of Essex and Venice, was one of the first projects dealing with the annotation of anaphora in Italian (Poesio et al., 2004). The VENEX corpus consists of articles from newspapers (Repubblica corpus) and task oriented dialogues of the Italian version of the Map Task corpus. The most relevant aspects of the coding scheme for our purposes are the solutions for problems of anaphoric annotation in Italian texts, like the annotation of clitics and empty subjects.

A more recent annotation effort is the I-CAB (Italian Content Annotation Bank) corpus (Magnini et al., 2006). I-CAB consists of articles from the regional newspaper l' Adige, annotated with a scheme close to that used for the annotation of the ACE-02 corpus (LDC, 2004).

3. Source of the Data

The LMC currently contains texts from two genres: texts from the Italian version of Wikipedia[2], and texts from blogs commented by the users and licensed under Creative Commons Attributions.

The Wikipedia texts cover a variety of topics mostly related to Trentino-Alto Adige/Südtirol, a region of North Italy. The texts refer to historical, geographical, and cultural items, like artworks, buildings, mountains, towns or prominent personalities.

Blog texts also refer to the region and include posts on a variety of topics followed by users' comments.

It is expected that the final version of the corpus will include newspaper texts from the regional newspaper l'Adige. The total amount of annotated data will consist of about 150,000 words for each genre.

4. Annotation methodology

4.1. Selection of the texts

One of the aims of our annotation effort is to provide annotated resources for Italian under the Creative Commons Attributions license. Consequently, for the annotation of blogs we look for texts published under this license or a similar one. Frequently, the owners of the blogs do not give any information about the conditions to share the content. This restricts substantially the amount of usable texts.

A second criterion used in the selection is the topic. We have selected material related to the Trentino/Süd Tirol region. The texts refer to entities situated in the region (like villages and towns, relevant people), events that happened in the region, or people and organizations that visited the region or have an influence on it (like artists that organize exhibitions in local museums, politicians, etc).

The last criterion is the length of the texts. Most of the chosen texts have a length of more than 400 words in order to be able to capture interesting anaphoric phenomena and large coreference chains. The application of this criterion for the Wikipedia texts was not very difficult. In the selection of texts of blogs the issue become harder, because very often one entry contains just a couple of sentences, or the advertisement for an event.

We decided to set the upper boundary for the length of the texts on the 2500 words, because it otherweise becomes difficult for the annotators to remember the entities that appear at the beginning of the text. Some articles of Wikipedia have more than 10.000 words, and for these cases we decided to select just fragments of the article, in which there are not pronouns referring to entities outside of the fragment. This constraint does not represent a problem for the selection of blogs, because the texts are shorter.

4.2. Extraction of the texts

The texts so far annotated for the LMC are taken from the Web. Web pages offer information in a visually structured way, structure that we keep as a part of the corpus. The preservation of layout information is especially relevant in the annotation of blog sites.

Blog pages contain text introduced by the author and comments introduced by users of the blog. Comments reflect opinions about the main post, about other comments, and update of information. They are written by different authors with different writing styles. In the comments we find mentions to entities that corefer with entities appearing in the main post, or in other comments. Another interesting feature is the mention of other comments using IDs or the name of the user.

[1] Available at `http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf`

[2] `http://it.wikipedia.org`

When using content from the web, subsequent usage might suffer from messy data – in our case, open or disguised advertising, and boilerplate are the main concerns. Consequently, the data need to be cleaned first. To this end, we use the KrdWrd tool (Steger and Stemle, 2009), a tool for the unified processing of web content. Here, we can use it to select the parts of a web page we want to keep and also add annotation categories in a consistent way.

4.3. Extraction of markables

The markables for anaphoric annotation are automatically extracted from raw texts through a newly developed pipeline. At the beginning, we use the tokenizer, POS tagger and sentence splitter from TextPro (Pianta et al., 2008). Then, we parse the data using the MALT dependency parser (Nivre et al., 2007) trained on the TUT treebank (Bosco et al., 2000). Afterwards, we use the produced dependency trees to create markables for all noun phrases (NPs).

The markables do not only contain the head of the NP, but all the pre- and postmodifiers, like adjectives, prepositional phrases or relative sentences that modify the head as we show in the example (1) (text wp_0133) (Translation: *"the first president of the Green group of the Parliament"*).

(1) [il primo presidente del gruppo parlamentare dei Verdi]
[del gruppo parlamentare dei Verdi]
[dei Verdi]

4.4. Human annotation

For the human annotation we use MMAX2 (Müller and Strube, 2006), which uses a stand-off format that allows marking and storing information in different levels.

After we import the output of our pre-processing pipeline in the format required by MMAX2, the last step before the annotation consists of the correction of possible errors in the markable boundaries and of the addition of not recognized markables. The main corrections that should be introduced are:

- As we explain in section 5.4. in our annotation there are discontinuous markables. We cannot extract this kind of markables from the parse tree, and the human annotator has to introduce them by hand.
- In Italian there are clitic pronouns attached to the verb (see Section 5.3.). In this case we select the verb and clitics as a markable, or as more than one if there is more than only one clitic attached. The annotator has to introduce this kind of markables by hand.
- Frequently the subject of the sentence is not phonetically realized. As we explain later in section 5.3., this non-phonetically realized pronouns are annotated as referring expressions, and they should be identified as markables. Since MMAX2 does not allow annotating empty markables, the annotator has to turn by hand the verb of the sentence into a markable.

Each markable is annotated with the tags of our annotation scheme presented in the next section.

5. Annotation scheme

The goal of the LMC annotation is to study anaphora and deictic reference. We use the term 'anaphora' to indicate context dependence, namely, the fact that many expressions in natural language derive at least part of their meaning from the meaning of expressions already introduced in the context. Anaphora thus defined covers a wide variety of phenomena. In this annotation, however, we are focusing on nominal anaphora, i.e., anaphoric reference using explicitly or implicitly realized noun phrases (NPs).

The LMC annotation scheme builds on the annotation scheme introduced in ARRAU (Poesio and Artstein, 2008) with significant modifications introduced to treat two linguistic phenomena of Italian, namely, empty subjects and incorporated clitics, and the use of discontinuous markables to annotate coordinations in specific cases.

Before illustrating in detail these novel features, we begin with a general description of the annotation scheme and its structure.

In general, all NPs, including definite, indefinite, pronominal, or proper names, are treated as markables. Each markable is annotated with a set of attributes and each attribute is associated with a set of values.

The attributes are of two broad categories: morphosyntactic and semantic.

The morphosyntactic attributes allow marking information about gender, number, and person. Each of these attributes is associated with a set of relevant values (e.g., `masculine` and `feminine` for **Gender**). In addition, a value `Underspecified` is used to mark NPs not encoding gender or number information, like *che* (*who*) in non restrictive relative clauses, and coordinations where the coordinated NPs are of different gender, number, or person.

The semantic annotation starts from a **Reference** attribute that can take the values `non-referring`, `discourse-new`, or `discourse-old` as described below.

5.1. Annotation of non-referring NPs

In the LMC, non-referring NPs are identified and marked with an appropriate value which specifies whether the NP is an expletive, an NP occurring in an idiomatic expression, a quantified NP, a coordination, or a predicate.

In case of quantified NPs in which the domain of quantification is anaphoric, the markable created for the domain of quantification is considered referential and annotated as explained in the next section.

The value `Coordination` is used to annotate coordinations like *John and Mary*. Each coordinated NP, *John* and *Mary*, however, is annotated as a referring expression and thus can serve as antecedent.

The value `Predicate` is used to mark predicative NPs. In the corpus, most predicative expressions appear in copular sentences and appositions. When these constructions involve proper names, as in 'Il Presidente della Repubblica Italiana, Giorgio Napolitano'(*The President of the Italian Republic, Giorgio Napolitano*), the markable for the proper name is treated as referential, while the markable for the whole NP is marked as a predicate. Another frequent example of predicative expressions in the LMC involves expressions like 'La città di Bolzano' *(The city of Bolzano)*. In this case, the whole expression is annotated as `predicate` while 'Bolzano' as a referring expression.

5.2. Annotation of referring NPs

Referring NPs are associated with two possible values, `new` and `old`, to indicate whether the NP is mentioned for the first time in the text or it has already been mentioned.

Both `new` and `old` markables are annotated with respect to their semantic category. The values for the category attribute combine the five ACE semantic classes with animacy and a concrete/abstract distinction, as illustrated below.

- `Person`: a single individual or a group of humans.
- `Organization`: corporations, agencies, and other groups of people defined by an established organizational structure.
- 'Geo-Political Entity' (`GSP`): geographical regions defined by political and/or social groups (e.g., a nation, its regions, etc.).
- `Location`: geographical entities such as geographical areas and landmasses, bodies of water, and geological formations. This value is also used for websites, telephone numbers, email addresses, etc.
- `Facility`: buildings and other permanent man-made structures and real estate improvements (e.g., houses, museums, airports, stations, roads, etc.).
- `Temporal`: dates and other temporal expressions.
- `Numerical`: percentages, units of measures, and expressions referring to money (e.g., 6 euros).
- `Animate`: not human animate objects.
- `Concrete`: not animate physical objects.
- `Abstract`: events, actions, plans, and abstract entities (e.g., the law).

Both `new` and `old` markables are also associated with a `related-object` attribute which indicates whether the NP stands in a bridging relation with a previously introduced NP. The two NPs are also linked by a pointer entered through the MMAX2 graphical interface. Annotators have been instructed to mark as related-object only three kinds of bridging relations: `part-of`, `set-member`, and `attribute`.

To annotate coreference, the attribute `old` is associated with a subset of attributes specifying several kinds of information about the antecedent expressions. These are:

- **Type of reference**, which allows identifying cases of discourse-deixis, although in our annotation no link is realized between the referring NP and the antecedent clause. The possible values for this attribute are `phrase-antecedent`, for NP antecedents, and `segment`, for discourse-deixis. When the `phrase-antecedent` value is selected, the NP is linked to its most recent antecedent through a pointer introduced using the MMAX2 graphical interface.
- **Phrase antecedent**, which allows marking plural reference. In these cases, all antecedents are identified through multiple pointers introduced through the MMAX2 graphical interface.
- **Ambiguity**, which allows identifying those cases in which the interpretation of the NP is ambiguous between two or more interpretations. The alternative antecedents are signaled through a distinct set of pointers entered through the MMAX2 graphical interface. In addition, the value `ambiguous-antecedent` allows specifying whether the antecedent for the markable was marked as ambiguous in the sense just described.

5.3. Annotation of clitics and empty-subjects

The LMC annotation scheme treats two significant linguistic phenomena of Italian: empty subjects and incorporated clitics. From the point of view of the annotation, these two phenomena are problematic since empty subjects are not at all realized in the surface form of the text and incorporated clitics are not realized as a separate word. We solved this problem by adopting a solution similar to that proposed in the VENEX annotation scheme – see (Poesio et al., 2004). Specifically, empty subjects are annotated by turning all the occurring finite verbal forms in the null-subject sentences into markables. To illustrate this, consider the following example taken from the corpus (wp_0012) (square brackets indicate a markable):

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

In this example, the subject of the second sentence is a morphologically null reference to the subject of the first sentence. In such cases, the verb at the beginning of the second sentence (*Nacque*) is turned into a markable and then annotated as if it contained the subject of the sentence, i.e. as a discourse-old markable referring to *Cesare Battisti*.

The solution adopted for incorporated clitics is similar, as illustrated in the following example taken again from wp_0012:

(3) ...[Il giudice] [gli] nego' [questa richiesta] e procedette invece ad acquistare [alcuni indumenti da [fargli] indossare]...
The judge to-him rejected this request and proceeded instead to buy some clothes to make-to-him wear.

In *fargli*, the clitic *gli* (*to him*) is attached to the verb *fare* (*make*). In such cases, the verbal form hosting the clitic is turned into a markable and then annotated as if it contained the clitic alone, i.e. as a discourse-old markable.
In order to specify whether a markable contains an NP or a verbal form standing for an empty subject or a clitic, we added the attribute **Markable type** with 'Nominal' and 'Verbal' as possible values. Verbal markables are associated with a **Verbal type** attribute that allows specifying whether the markable is a clitic or an empty subject. As already mentioned, these markables are then annotated with the set of attributes previously illustrated for nominal markables.

5.4. Coordination and discontinuous markables

The need to use discontinuous markables arises for cases of coordinations of modified NPs where the semantic unity of one NP is interrupted, as in *the green and red cars*. The LMC presents a variety of cases of coordinations requiring discontinuous markables. One example is illustrated in (4) (wp_0004) where, in order to preserve semantic unity and annotate both elements of the coordination (*Enrico Conci* and *Elsa Conci*), *Enrico Conci* is realized as a discontinous markable.

(4) ...[[Enrico] ed [Elsa [Conci]]....

6. Description of the corpus

6.1. Wikipedia dataset

The current set of annotated texts of Wikipedia consists of 144 files, with 142 K of words. In this dataset we have selected 44.5 K of markables for the annotation.

Of the total set of markables, 0.5% are discontinuous markables, and the same quantity (0.5%) are clitics attached to the verb. 4.5% of the anaphoric expressions are empty subjects, and all of them are linked to a previous antecedent. The resolution of this kind of anaphora is an interesting topic because they are part of the coreference chains.

57.8% of the markables have been tagged as `discourse-new`, 28.5% of the markables as `discourse-old` and 13.7% of the markables have been classified as `non-referring` expressions.

More than 50% of the `non-referring` markables of this dataset have been tagged as cases of predication, and 34% as cases of coordination. The distribution of categories can be seen on Fig. 1

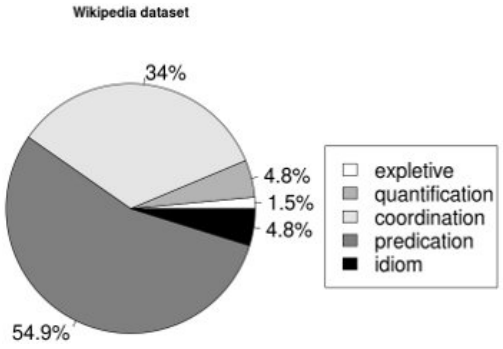

Figure 1: Distribution of the non-referring expressions in the Wikipedia dataset

6.2. Blogs dataset

The current set of annotated blog texts consist of 66 files with 50K words and 15K of markables for the annotation.

On the total set of markables, 0.8% are discontinuous markables, an 1% clitics attached to the verb. 4.7% of the anaphoras are empty subjects, and all of them are linked to a previous antecedent.

64.7% of the markables have been tagged as `discourse-new`, 23.6% of the markables as `discourse-old` and 11.7% of the markables have been classified as `non-referring` expressions.

Most of the referring expressions are cases of predication or coordination with a high presence of occurrences of markables tagged with one of the other categories, especially expletives (see Fig. 2).

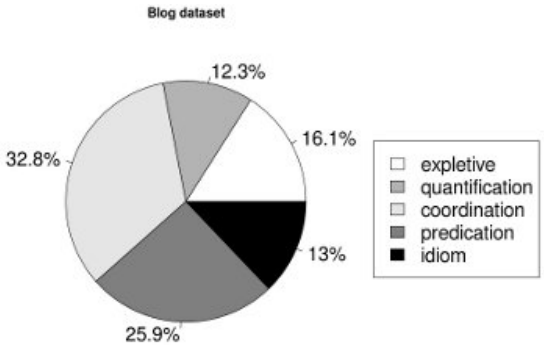

Figure 2: Distribution of the non-referring expressions in the blogs dataset

6.3. Main differences

The main difference between both datasets is the distribution of the categories of non-referring expressions. Fig. 1 shows that most of the non referring expressions of the Wikipedia dataset are predicates. Only 25% of the non-referring expressions of the blogs dataset belong to this category. The main reason seems to be the necessity of precision and more accurate descriptions of entities and events of the encyclopedic texts.

The percentage of idiomatic expressions found in the blogs dataset is 3 times higher than in the Wikipedia dataset. The reason is that often the style is less formal and more conversational, above all in the commentaries posted by the users.

Another important difference is the higher amount of expletives and clitics in the blogs dataset.

7. Inter-annotator agreement

We have tested the reliability of our annotation scheme carrying out separate agreement studies for several features of the annotation scheme. For these studies we have used the Kappa coefficient (Carletta, 1996) between two annotators.

The first set of studies measures the **reliability of the annotation of features on the markable level**. These features are the information status, the referentiality of the markable and the semantic type of referring expressions.

- **Information status.** The possible values for this attribute are `discourse-old`, `discourse-new` and `non-referring`. The value of Kappa is κ = 0.80.
  We have observed in the confusion matrix that the most common disagreement is between the values `new` and `non_referring`.

- **Basic annotation of the markable.** That is the annotation performed before the annotators begin with the annotation of anaphora. The possible values are `discourse-new`, `segment-antecedent` (for discourse deixis), `phrase-antecedent`, and for non referring NPs `expletive`, `quantifier`, `predicate`, `coordination` and `idiom`. The Kappa value is κ = 0.79.
  The most common disagreement is between the tags `discourse-new` and `predicate`.

- **Semantic type.** The value of Kappa for this feature is κ= 0.85.
  The most common disagreements observed in the confusion matrix are between the categories abstract and concrete and between the categories gps and organization.

The second set of studies measures the **reliability of the annotated anaphoric links**. First of all we have carried out a study for the selection of antecedent of all anaphoric links annotated in the experiment.

As we have mentioned earlier, empty pronouns in the subject position of the sentence and clitics attached to the verb are two phenomena that appear with a relevant frequency in Italian texts. We have carried out separate studies to test the reliability of the annotation scheme for these phenomena.

- **Link to the antecedent:** The value of Kappa for the annotation of links from markables tagged as `old` to the immediate antecedent is κ = 0.88.

- **Antecedent of clitics:** The value of Kappa for the annotation of links from markables realized as incorporated clitics to the immediate antecedent. κ = 0.84.

- **Antecedent of empty pronouns.** The value of Kappa for the annotation of links from empty pronouns to the immediate antecedent. κ = 0.93.

8. Conclusions and further work

The work presented in this article is aimed at contributing to two significant issues for CL research: the lack of annotated anaphoric resources for Italian and the increasing interest for social Web and the texts produced by it.

Markables have been annotated with morphosyntactic and semantic information and anaphoric links. The annotation scheme takes into account two important phenomena of the Italian language, the clitics attached to the verb and the empty subjects. Reliability studies for the annotation of the mentioned phenomena, and for the annotation of anaphoric links in general show a good agreement between annotators. These results indicate that the scheme can be re-used to annotate other data sets.

A part of the LMC, the Wikipedia dataset, will be used for the SEMEVAL 2010 multilingual coreference evaluation[3].

In the near future we will introduce a new annotation level for the blog dataset that captures information about the layout, and we will extend the annotation to a new set of news paper articles.

We plan to distribute the corpus in two formats: the MMAX2 standoff format and a tabular format inspired by the 2008/2009 CoNLL shared tasks on syntactic and semantic dependencies[4] and used for the SEMEVAL 2010.

The Wikipedia and Blogs datasets will be distributed under Creative Commons Attributions lincense.

9. References

A. Balahur, E. Boldrini, A. Montoyo, and P. Martínez-Barco. 2009a. Opinion and generic question answering systems: a performance analysis. In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 157–160, Morristown, NJ, USA. Association for Computational Linguistics.

A. Balahur, E. Lloret, E. Boldrini, A. Montoyo, M. Palomar, and P. Martinez-Barco. 2009b. Summarizing threads in blogs using opinion polarity. In Constantin Orăsan, Laura Hasler, and Corina Forascu, editors, *Proceedings of the International Workshop on Events in Emerging Text Types (eETTs)*, pages 5 – 13, Borovets, Bulgaria.

P. Bo and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrival*, 2(1 & 2):1 – 135.

[3] `http://stel.ub.edu/semeval2010-coref`
[4] `http://ufal.mff.cuni.cz/conll2009-st/`

Ester Boldrini, Alexandra Balahur, Patricio Martnez-Barco, and Andrs Montoyo. 2009. Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In Robert Stahlbock, Sven F. Crone, and Stefan Lessmann, editors, *DMIN*, pages 491–497. CSREA Press.

C. Bosco, V. Lombardo, D. Vassallo, and L. Lesmo. 2000. Building a treebank for italian: a data-driven annotation scheme. In *Proceedings of the LREC-00*, Athens. Greece.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

I. Hendrickx and V. Hoste. 2009. Coreference resolution on blogs and commented news. In *DAARC '09: Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium on Anaphora Processing and Applications*, pages 43–53, Berlin, Heidelberg. Springer-Verlag.

D.J. Hopkins and G. King. 2010. A method of automated nonparametric contente analysis for social science. *American Journal of Political Science*, 54(1):229 – 247.

Linguistic Data Consortium LDC, 2004. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, version 5.6.1*.

B. Magnini and M. Poesio. 2009. Content extraction meets the social web in the livememories project. In *Proceedings of the AT4DL*.

B. Magnini, E. Pianta, Ch. Girardi, M. Negri, L. Romano, M. Speranza, and R. Sprugnoli. 2006. I-cab: the italian content annotation bank. In *Proceedings of the LREC-06*, Genova, Italia.

R. Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings of the NAACL*.

Ch. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with mmax2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt a. M., Germany.

N. Nicolov, F. Salvetti, and S. Ivanova. 2008. Sentiment analysis: Does coreference matter? In *Proceedings of the Symposium on Affective Lenguage in Human and Machine*, Aberden. UK.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kbler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 2(13).

E. Pianta, Ch. Girardi, and R. Zanoli. 2008. The textpro tool suite. In *Proceedings of LREC-08*.

M. Poesio and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of LREC-08*, Marrakech, Marocco.

M. Poesio, R. Delmonte, A. Bristot, L. Chiran, and S. Tonelli. 2004. The venex corpus of anaphora and deixis in spoken and written italian. Manuscript.

S. Ponzetto and M. Strube. 2007. Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30.

J. M. Steger and E. W. Stemle. 2009. The architecture for unified processing of web content. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, Donostia-San Sebastian. Basque Country.