Towards a Balanced Named Entity Corpus for Dutch

Bart Desmet[1,2] **and Véronique Hoste**[1,2]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
bart.desmet, veronique.hoste@hogent.be
[2]Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium

Abstract

1. Introduction

Named Entity Recognition (NER) is the task of automatically identifying names within text and classifying them into categories, such as persons, locations and organizations. NER started as an information extraction subtask, but has since evolved into a distinct task essential for information retrieval, question answering, and as a preprocessing step for coreference resolution and various other problems. An extensive literature on the subject exists (see for example (Tjong Kim Sang, 2002), (Chinchor, 1998)), with NER approaches roughly falling into three categories: hand-crafted, machine learning and hybrid systems. Hand-crafted approaches use gazetteer lists and require manual rule creation, a time-consuming process which hinders easy porting to new domains or languages. Supervised machine learning solutions, on the other hand, rely on an annotated training corpus to infer patterns associated with named entities, based on orthographic, syntactic, lexical and contextual features. Hybrid systems combine both approaches. Such systems are in widespread use and have proven their effectiveness, with (Zhou and Su, 2002) reporting near-human performance on English data.

The bottleneck for the development of machine learning applications is its dependence on, preferably large, annotated training corpora. Named entity resources for English include the manually annotated data sets from the MUC-7 Named Entity Task ((Chinchor, 1998), 162,692 tokens), the CoNLL-2003 shared task ((Tjong Kim Sang and De Meulder, 2003), 301,418 tokens) and the BBN Pronoun Coreference and Entity Type Corpus ((Weischedel and Brunstein, 2005), 1,173,766 tokens). For Dutch, however, the data from the CoNLL-2002 shared task ((Tjong Kim Sang, 2002), 309,686 tokens from four editions of the Belgian newspaper "De Morgen" of 2000) constitute the only corpus annotated with named entity information that is readily available at present.

The pressing need for a substantial corpus of Dutch text, not only for NER, is addressed in the STEVIN[1]-funded SoNaR project[2]. It aims to produce a 500-million-word reference corpus of written Dutch containing a wide spectrum of genres and text types (Oostdijk et al., 2008), including a 1-million-word subset with a number of manually corrected annotation layers, including four semantic ones: named entities, coreference relations, semantic roles and spatiotemporal expressions (see Schuurman et al., (2009)). The subset contains the various text types, reflecting the global corpus design. This diversity, which was particularly lacking in the Dutch CoNLL-2002 data set, should allow for a more robust classifier and better cross-corpus performance.

For the named entity annotation of the corpus, we developed new annotation guidelines, based on the guidelines from MUC-7 (Chinchor and Robinson, 1997) and ACE (LDC, 2008). A number of adaptations were made, most notably the addition of separate classes for products and events, and the annotation of metonymy.

In the remainder of this abstract, we will discuss and motivate the annotation guidelines in Section 2, we present an evaluation of the guidelines based on inter-annotator agreement scores in Section 3 and give an overview of the use of the guidelines within the context of the SoNaR project in Section 4. Section 5 concludes this abstract.

2. Annotation guidelines

The SoNaR named entity annotation guidelines[3] are based on the MUC-7 and ACE annotation schemes for English named entities. Annotation of numerical and temporal expressions was considered beyond the scope of the NE task. The aim of the new guidelines was to achieve consistent and fine-grained annotation of Dutch text. To that end, the guidelines describe the delimitation of named entities (see 2.1.), the classification into main types (2.2.) and subtypes (2.3.), and the markup of metonymic usage (2.4.). For an overview of the possible annotations, see Figure 1.

2.1. Span

Named entities are often defined as "unique identifiers of referents in reality". In practice, it is often unclear whether a given phrase should be considered a unique identifier or not. This is especially true for named entities of the types

[1]http://taalunieversum.org/taal/technologie/stevin/
[2]http://lands.let.ru.nl/projects/SoNaR/
[3]http://lt3.hogent.be/sonar/share/AnnotatierichtlijnenNE20091019.pdf, in Dutch

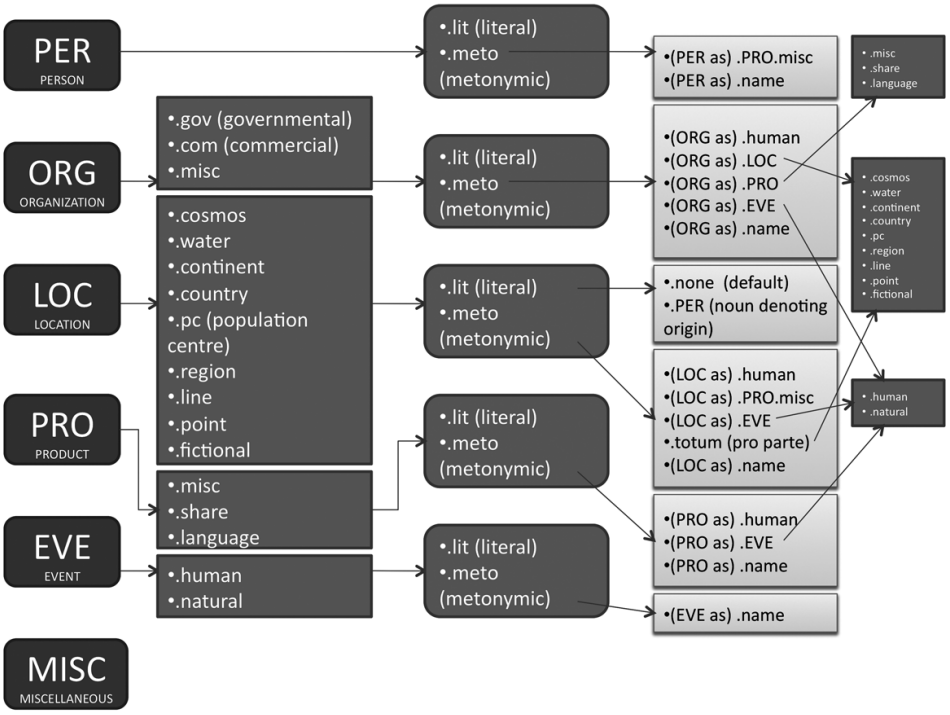

Figure 1: Annotation scheme for named entities, with categories for main type, subtype, usage and metonymic roles.

product, event and miscellaneous (see 2.2.). Consider the following example:

(1) Koning Albert zal de twee Koninklijke Besluiten van minister van Werk Peter Vanvelthoven (sp.a) niet ondertekenen.
English: King Albert will not sign the two Royal Decrees by the minister of Employment, Peter Vanvelthoven (sp.a).

In sentence 1, it is debatable whether "King", "Royal Decrees" and "Employment" should be considered (part of) a unique identifier. For reasons of consistency, a pragmatic approach was taken to the delimitation of named entities. All words starting with a capital letter that are not the first word of a sentence are taken to be named entities. All sentence-initial or uncapitalized words that can unequivocally be considered unique identifiers are annotated as well. Sentence 1 will be annotated as follows:

(2) Koning [Albert] zal de twee [Koninklijke Besluiten] van minister van [Werk] [Peter Vanvelthoven] ([sp.a]) niet ondertekenen.
English: King [Albert] will not sign the two [Royal Decrees] by the minister of [Employment], [Peter Vanvelthoven] ([sp.a]). (sp.a is a Belgian political party)

Named entities can be part of a word that as a whole is not a named entity, e.g. "London-based". In English, such structures are rare and will often be annotated fully as MISC, or not at all (Nothman et al., 2009). Given the frequency of concatenated compounds in Dutch, we chose to annotate named entities word-internally:

(3) [Apple]topman [Steve Jobs] kondigde het [iPhone]platform op [Macworld 2007] aan.
English: [Apple] [CEO] [Steve Jobs] announced the [iPhone] platform at [Macworld 2007].

2.2. Main types

Tokens marked as named entities can be classified as one of six main types, namely person (PER), organization (ORG), location (LOC), product (PRO), event (EVE) and miscellaneous (MISC). PER, ORG and LOC are the usual suspects in named entity annotation, with MISC sometimes acting as a backup class. We added the PRO and EVE categories to obtain good coverage of possible named entities and to allow for consistent metonymy roles (see 2.4.). The MISC category was reserved for instances produced by the broad definition of span (2.1.) that fitted more than one or none of the five other main types.

2.3. Subtypes

For the ORG, LOC, PRO and EVE classes, mutually exclusive subtypes are to be annotated. The motivation for subtypes was twofold:

1. It allows for fine-grained annotation, as required for question answering tasks, without compromising a robust, coarse-grained main type structure.

2. It provides useful information for the classification of usage (2.4.) and for the other semantic annotation layers in the SoNaR project.

The markables in Sentence 3 would be classified as follows:

[Apple]$_{ORG.com}$topman [Steve Jobs]$_{PER}$ kondigde het [iPhone]$_{PRO.misc}$platform op [Macworld

2007]$_{EVE.human}$ aan.
English: [Apple]$_{ORG.com}$ [CEO]$_{MISC}$ [Steve Jobs]$_{PER}$ announced the [iPhone]$_{PRO.misc}$ platform at [Macworld 2007]$_{EVE.human}$.

2.4. Usage

An important issue we wanted to address in our annotation scheme was the metonymic use of named entities. Consider Sentence 4:

(4) Het [Witte Huis] koos voor moderne werken, waaronder een [Rothko].
English: The [White House] opted for modern works of art, including a [Rothko].

Cases like "White House" being classified as LOC rather than ORG are a common mistake (Nothman et al., 2009). By marking whether a NE is used literally or metonymically, we can consistently label named entities for their literal main type, and use metonymic roles to point to their intended main type (PER, ORG, LOC, PRO or EVE). This approach was inspired by Markert and Nissim (2002) and Markert and Nissim (2007). Because it is often impracticable to determine whether a NE is used metonymically as PER or as ORG, we combined them in the intended type "human" (see for example 5, where "White House" might refer to a PER, namely the U.S. President, or to an ORG-like group of people such as the White House staff). When a name is used as a mere signifier, the intended type is "name".

(5) Het [Witte Huis]$_{LOC.point.meto.human}$ koos voor moderne werken, waaronder een [Rothko]$_{PER.meto.PRO.misc}$.
English: The [White House]$_{LOC.point.meto.human}$ opted for modern works of art, including a [Rothko]$_{PER.meto.PRO.misc}$.

Marking metonymy does not only do away with confusable main types, it should also benefit the automatic annotation of other semantic layers. For example, a coreferential resolution algorithm could link an inanimate noun phrase like "the painting" to "Rothko" in Sentence 5 if it has access to NE classifier output that does not only mark "Rothko" literally as an (animate) person, but also metonymically as a product.

3. Guideline evaluation

In order to evaluate the guidelines, two linguists annotated a set of eight randomly selected texts from the corpus, containing 14,244 tokens in total. Two evaluation metrics were used: Kappa (Carletta, 1996) and F-score (β = 1) (Van Rijsbergen, 1979). F-scores were calculated by taking one annotator as the gold standard and scoring the annotations of the other for precision and recall. This yields the same results as averaging the precision scores of both annotators, when using the other as a gold standard.

Scores were calculated on 5 levels: span, main type, subtype, usage and metonymic role. For each level, scores were calculated on the entire set, and on a subset containing only those tokens (i) on which both annotators agreed on the preceding level, and (ii) which bore annotation on the current level (MISC and PER, for example, are not included in the subset for subtype). The results can be found in Table 1, in which absolute counts for span, main type and usage are also included.

The results show high agreement scores for all levels, most notably span. However, both global metrics fail to indicate low agreement for minority classes, as is the case for metonymic usage. F-scores per class have been calculated to detect problematic classifications and to refine the guidelines accordingly.

4. Annotation implementation

The SoNaR corpus will comprise a wide variety of texts, including newswire, manuals, autocues, online material, fiction and reports, for a total of 500 million words. A representatively diverse 1-million-word subset is being annotated manually, and serve as the gold standard for the automatic annotation of the entire corpus. This diversity is essential to training automatic classifiers that can be applied on the corpus as a whole, and should also make it an interesting corpus for research on domain adaptation (Nothman et al., 2009).

Manual annotation is done using the MMAX2 annotation tool[4]. Each annotation layer is stored as one or several standoff XML files. For the named entity task, six XML files are created - one per main type. Annotation speed averages around 4,000 words per hour. Taking into account the verification of the annotations by a second annotator, the actual annotation speed will be closer to 2,500 words per hour.

5. Conclusion and future work

This abstract presented the named entity annotation guidelines for the SoNaR project. Their aim is to provide consistent and fine-grained annotations that capture useful information for subsequent classification tasks. To this end, a pragmatic approach was taken for the delimitation of named entities, resulting in high inter-annotator agreement scores for span.

In the full version of this paper, we intend to provide statistics on the distribution of the various types, usage and metonymic roles, and discuss the generalization performance of classifiers trained on the fully annotated corpus.

6. References

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

N. Chinchor and P. Robinson. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.

N. Chinchor. 1998. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*.

[4]http://www.eml-research.de/english/research/nlp/download/mmax.php

Level	Total set		Subset			
	Kappa	**F**$_{\beta=1}$	**Kappa**	**F**$_{\beta=1}$	**Tokens**	**Distribution**
Span	0.97	99.62	0.97	99.62	14244	13293 non-NE, 897 NE, 54 NA
Main type	0.94	99.23	0.92	93.76	897	150 PER, 225 ORG, 241 LOC, 115 PRO, 62 EVE, 48 MISC, 56 NA
Subtype	0.92	99.12	0.94	97.67	643	32 NA
Usage	0.91	98.93	0.93	94.58	793	733 literal, 17 metonymic, 43 NA
Role	0.91	98.90	1.00	100.00	17	0 NA

Table 1: Inter-annotator agreement scores per level, token count and distribution (NA = no agreement).

LDC, 2008. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6*. Linguistic Data Consortium, Philadelphia. http://projects.ldc.upenn.edu/ace/.

K. Markert and M. Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. In *Proceedings of the Third International Language Resources and Evaluation (LREC'02)*, pages 1385–1392, Las Palmas, Spain.

K. Markert and M. Nissim. 2007. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague, Czech Republic.

J. Nothman, T. Murphy, and J.R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 612–620, Athens, Greece.

N. Oostdijk, M. Reynaert, P. Monachesi, G. Van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

I. Schuurman, V. Hoste, and P. Monachesi. 2009. Cultivating trees: Adding several semantic layers to the Lassy treebank in SoNaR. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, The Netherlands.

E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147, Edmonton, Canada.

E.F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 155–158, Taipei, Taiwan.

C.J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth, London.

R. Weischedel and A. Brunstein, 2005. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia, USA.

G.D. Zhou and J. Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480, Philadelphia.