Latvian and Lithuanian Named Entity Recognition with TildeNER

Mārcis Pinnis

Tilde
75a Vienibas gatve, LV-1004, Riga, Latvia
marcis.pinnis@tilde.lv

University of Latvia
19 Raina Blvd., LV-1586, Riga, Latvia
marcis.pinnis@lais.lv

Abstract

In this paper the author presents TildeNER – an open source freely available named entity recognition toolkit and the first multi-class named entity recognition system for Latvian and Lithuanian languages. The system is built upon a supervised conditional random field classifier and features heuristic and statistical refinement methods that improve supervised classification, thus boosting the overall system's performance. The toolkit provides means for named entity recognition model bootstrapping, plaintext document and also pre-processed (morpho-syntactically tagged) tab-separated document named entity tagging and evaluation on test data. The paper presents the design of the system, describes the most important data formats and briefly discusses extension possibilities to different languages. It also gives evaluation on human annotated gold standard test corpora for Latvian and Lithuanian languages as well as comparative performance analysis to a state-of-the art English named entity recognition system using parallel and strongly comparable corpora. The author gives analysis of the Latvian and Lithuanian named entity tagged corpora annotation process and the created named entity annotated corpora.

Keywords: named entity recognition, Latvian and Lithuanian languages, bootstrapping

1. Introduction

Named entity recognition (NER) has been actively researched for over 20 years. Most of the research has, however, been focussed on resource rich languages, for instance, English French and Spanish. The scope of this paper covers the task of named entity recognition for two under-resourced languages – Latvian and Lithuanian. The author presents an open source freely available toolkit named *TildeNER* that makes use of existing supervised learning methodology (for instance, the Stanford NER conditional random field classifier (Finkel et al., 2005)) enriched with heuristic refinement methods in order to bootstrap NER models using unlabelled data, thus, creating a "*highly supervised*" semi-supervised named entity recognizer.

Latvian and Lithuanian are the state languages of two European Union member countries - Latvia and Lithuania. Both languages feature rich morphology with high morphological ambiguity and a relatively free order of constituents in sentences, thus, making the task of named entity recognition more difficult than, for instance, for English.

The current dominant approach to developing named entity recognition systems is supervised learning (Nadeau and Sekine, 2007). This, however, means that a prerequisite for NER model training is a large named entity (NE) annotated data corpus. For resource rich languages this is not an issue, but for under-resourced languages (for instance, the Baltic languages) is. For Latvian and Lithuanian there has been very little previous research in the field of named entity recognition. Most of the existing research has dealt with only toponym recognition, for instance, Skadiņa (2009) describes toponym recognition from image annotations using lexicons and patterns. Also the lack of annotated named entity corpora for both languages does not allow (without significant financial input for corpora creation) the development of a truly supervised NER system. Because of the available resource constraints, for Latvian and Lithuanian a semi-supervised NER system development approach was selected, more precisely, bootstrapping. The systems presented in the paper are, therefore, the first multi-class NER tools created for Latvian and Lithuanian.

The main reason for the development of the Latvian and Lithuanian NER systems has been to tag NEs in comparable corpora for further bilingual NE alignment using NE mapping methods in the ACCURAT project[1]. It is also planned to use the NER systems as a pre-processing step in machine translation in order to create NE-aware translations.

The next chapter gives a description of the NE-annotated corpora followed by a section on the design and methods applied in *TildeNER* and evaluation in section four. The paper is finalized with conclusions and a discussion of future work.

2. Annotated Corpora

For the task of named entity recognition relatively small NE annotated corpora was created. The corpora for both languages consists of IT localization (software reviews, manuals and other IT related articles), news (current news from news web portals) and Wikipedia articles in equal proportions. The first two parts were acquired using comparable corpora web crawling tools developed within the ACCURAT project[2]. The corpora statistics is shown in Table 1.

For the annotation task, NE mark-up guidelines[3] were prepared. The guidelines are mostly compliant with the MUC-7 (Chinchor, 1998) NE annotation guidelines (adaptation to Latvian and Lithuanian was performed as

[1] Report on information extraction from comparable corpora,
[2] Tools for building comparable corpus from the Web, public deliverable of the project ACCURAT, 2011.
[3] Published as part of TildeNER in the „Toolkit for multi-level alignment and information extraction from comparable corpora", public deliverable of the project ACCURAT, 2011.

well as minor contradictions were resolved). The following NE categories were annotated: organization, person name, location, product, date, time, money.

	Latvian	Lithuanian
Document count		
Seed	40	37
Development	25	33
Test	66	55
Total	**131**	**125**
Word count		
Seed	20 959	18 852
Development	10 053	17 827
Test	41 208	36 239
Total	**72 220**	**72 918**

Table 1: Latvian and Lithuanian corpora statistics.

The corpora were annotated by two annotators and disagreements were resolved by a third annotator for both languages. The inter-annotator agreement between the first two annotators using the Cohen's Kappa statistic (Cohen, 1968) is 0.885 for Latvian and 0.822 for Lithuanian. This score, however, represents the overall complexity of the corpora including non-entities strictly classified as non-entities by both annotators. This score does not represent the actual NE annotation complexity and difficulties in NE border detection; that is, adding or removing non-entity data (tokens/sentences) will result in respectively higher or lower inter-annotator agreement. Therefore, separate NE category and NE border detection inter-annotator agreement scores are given in Table 2. The token level agreement scores do not consider cases where both annotators annotated a token as a non-entity.

	Latvian	Lithuanian
Full NE agreement		
NE border agreement	0.749	0.671
Category agreement on matching borders	0.964	0.967
Token level agreement		
LOCATION	0.790	0.703
ORGANIZATION	0.708	0.623
PERSON	0.932	0.910
PRODUCT	0.641	0.683
DATE	0.812	0.696
TIME	0.713	0.662
MONEY	0.785	0.599
Total token agreement	**0.807**	**0.723**

Table 2: Inter-annotator agreement on Latvian and Lithuanian corpora.

In the process of annotation a tool named *NESimpleAnnotator* was used (released together with *TildeNER*). The annotation tool allows fast one-dimensional (non-hierarchical) annotation of NEs of the defined categories. The annotation tool also features disambiguation functionality for a judge. The annotation tool in the disambiguation view is shown in Figure 1.

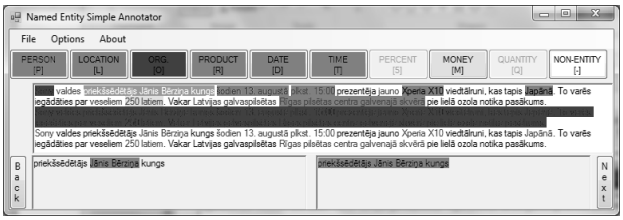

Figure 1: Disambiguation view of *NESimpleAnnotator*

After annotation both corpora were split in seed, development and test sets. The development set is used in refinement method parameter tuning and feature function selection processes and the test set is used for final evaluation. The NE statistics in the disambiguated corpora is shown in Table 3 for both Latvian and Lithuanian.

NE Type	Seed	Development	Test
Latvian			
DATE	498	249	843
LOCATION	682	479	1 453
MONEY	123	18	148
ORGANIZATION	464	219	966
PERSON	267	172	601
PRODUCT	381	103	382
TIME	200	46	107
Total	**2 615**	**1 286**	**4 500**
Lithuanian			
DATE	548	297	711
LOCATION	470	563	1 086
MONEY	150	147	313
ORGANIZATION	240	275	603
PERSON	202	169	604
PRODUCT	174	310	389
TIME	67	57	109
Total	**1 851**	**1 818**	**3 815**

Table 3: Latvian and Lithuanian NE annotated corpora statistics.

The NE annotated data is stored in plaintext format containing MUC-7 style NE tags. A format sample is given in Figure 2. This format is also used when *TildeNER* performs automatic NER on user provided plaintext documents.