Named Entity Recognition in Estonian

Alexander Tkachenko
Institute of Computer Science
University of Tartu
Liivi 2, Tartu, Estonia
alex.tk.fb@gmail.com

Timo Petmanson
Institute of Computer Science
University of Tartu
Liivi 2, Tartu, Estonia
timo_p@ut.ee

Sven Laur
Institute of Computer Science
University of Tartu
Liivi 2, Tartu, Estonia
swen@math.ut.ee

Abstract

The task of Named Entity Recognition (NER) is to identify in text predefined units of information such as person names, organizations and locations. In this work, we address the problem of NER in Estonian using supervised learning approach. We explore common issues related to building a NER system such as the usage of language-agnostic and language-specific features, the representation of named entity tags, the required corpus size and the need for linguistic tools. For system training and evaluation purposes, we create a gold standard NER corpus. On this corpus, our CRF-based system achieves an overall F_1-score of 87%.

1 Introduction

Named Entity Recognition (NER) is the task of identification of information units in text such as person names, organizations and locations. It is an important subtask in many *natural language processing* (NLP) applications such as text summarization, information filtering, relation extraction and question answering. NER has been extensively studied for widely spoken languages such as English with the state-of-the-art systems achieving near-human performance (Marsh and Perzanowski, 1998), but no research has yet been done in regards to Estonian.

The main difference of Estonian, a Finno-Ugric language, compared to English is high morphological richness. Estonian is a synthetic language and has relatively high morpheme-per-word ratio. It has both agglutinative and fusional (inflective) elements: morphemes can express one or more syntactic categories of the word. Although Estonian is considered a subject-verb-object (SVO) language, all phrase permutations are legal and widely used. These factors make NLP for Estonian particularly complicated.

In this work, we address the problem of NER in Estonian using supervised learning approach. We explore common issues related to building a NER system such as the usage of language-agnostic and language-specific features, the representation of named entity tags, the required corpus size and the need for linguistic tools.

To train and evaluate our system, we have created a gold standard NER corpus of Estonian news stories, in which we manually annotated occurrences of locations, persons and organizations. Our system, based on Conditional Random Fields, achieves an overall cross-validation F_1-score of 87%, which is compatible with results reported for similar languages.

Related work. The concept of NER originated in the 1990s in the course of the Message Understanding Conferences (Grishman and Sundheim, 1996), and since then there has been a steady increase in research boosted by evaluation programs such as CoNLL (Tjong Kim Sang and De Meulder, 2003) and ACE (ACE, 2005). The earliest works mainly involved using hand-crafted linguistic rules (Grishman, 1995; Wakao et al., 1996). Rule-based systems typically achieve high precision, but suffer low coverage, are laborious to build and and not easily portable to new text domains (Lin et al., 2003). The current dominant approach for addressing NER problem is supervised machine learning (Tjong Kim Sang and De Meulder, 2003). Such systems generally read a large annotated corpus and induce disambiguation rules based on discriminative features. Frequently used techniques include Hidden Markov Models (Bikel et al., 1997), Maximum Entropy Models (Bender et al., 2003) and Linear Chain Conditional Random Fields (McCallum and Li, 2003). The downside of supervised learning is the need for a large,

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 78–83,
Sofia, Bulgaria, 8-9 August 2013.

annotated training corpus.

Recently, some research has been done on NER for highly inflective and morphologically rich languages similar to Estonian. Varga and Simon (2007) report F_1-score of 95% for Hungarian in business news domain using a Maximum Entropy classifier. Notably, authors state that morphological preprocessing only slightly improves the overall performance. Konkol and Konopík (2011) also use Maximum Entropy based approach for NER in Czech achieving 79% F_1-score. Pinnis (2012) reports F-score of 60% and 65% for Latvian and Lithuanian languages respectively using CRF classifier with morphological preprocessing and some custom refinements. Küçük and others (2009) describe a rule-based NER system for Turkish language which achieves F_1-score of 79%. We observe that the reported results are notably inferior compared to well-studied languages such as English. This can be explained by the language complexity and the lack of required linguistic tools and annotated corpora.

2 The Corpus

Papers on NER for English language commonly use publicly available named entity tagged corpora for system development and evaluation (Tjong Kim Sang and De Meulder, 2003; Chinchor, 1998). As no such resources are available for the Estonian, we have built our corpus from scratch. Our corpus consists of 572 news stories published in the local online newspapers Delfi[1] and Postimees[2] between 1997 and 2009. Selected articles cover both local and international news on a range of topics including politics, economics and sports. The total size of the corpus is 184,638 tokens.

The raw text was preprocessed using the morphological disambiguator `t3mesta` (Kaalep and Vaino, 1998). The processing steps involve tokenization, lemmatization, part-of-speech tagging, grammatical and morphological analysis. The resulting dataset was then manually name entity tagged. Due to the limited resources, the corpus was first tagged by one of the authors and then examined by the other, after which conflicting cases were resolved. Following the MUC guidelines (Chinchor, 1998), we distinguish three types of entities: person names (PER), locations (LOC) and organizations (ORG). Words that do not fall into any of these categories were tagged as *other* (O). We assume that named entities do not overlap. In case a named entity is contained within another named entity, only the top-level entity is annotated. Table 1 and Figure 1 give an overview of named entity occurrences in the corpus.

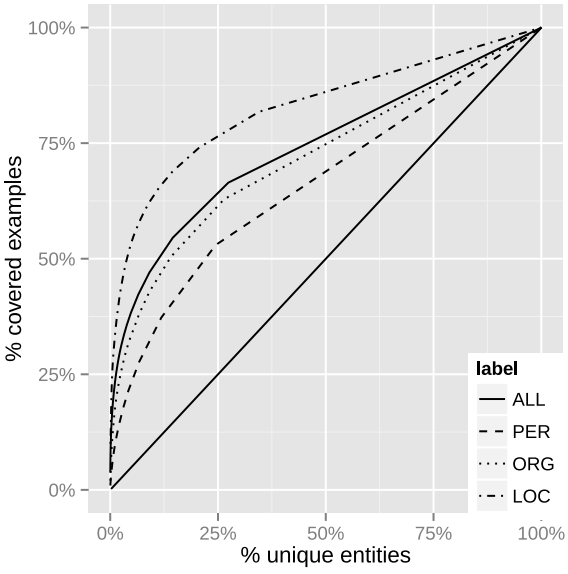

Figure 1: Cumulative number of examples covered by unique entities, starting with the most frequent.

	PER	LOC	ORG	Total
All	5762	5711	3938	15411
Unique	3588	1589	1987	7164

Table 1: Number of named entities in the corpus.

The corpus is organized closely following CoNLL03 formatting conventions (Tjong Kim Sang and De Meulder, 2003). In a data file, each line corresponds to a word with empty lines representing sentence boundaries. Each line contains four fields: the word itself, its lemma, its grammatical attributes[3] and its named entity tag. Named entity tags are encoded using a widely accepted BIO annotation scheme (Ramshaw and Marcus, 1995). Figure 2 demonstrates an example sentence.

The corpus is freely available for research purposes and is accessible at the repository of public language resources of Tartu University (Laur et al.,

[1] `http://delfi.ee`

[2] `http://postimees.ee`

[3] Definition of the attributes can be found at `http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=en`

11.	11.+0	_O_ ?	O
juunil	juuni+l	_S_ sg ad	O
laastas	laasta+s	_V_ s	O
tromb	tromb+0	_S_ sg n	O
Raplamaal	Rapla_maa+l	_H_ sg ad	B-LOC
Lõpemetsa	Lõpe_metsa+0	_H_ sg g	B-LOC
küla	küla+0	_S_ sg n	I-LOC
.	.	_Z_	O

Figure 2: An example sentence in the corpus: On the 11th of June, a tornado devastated Lypemetsa village in Rapla county.

2013).

3 System Overview

Two important components in the design of a NER system are features and a learning algorithm. Features encode characteristic attributes of words relevant for the classification task. Possible examples of features are word lemma, part of speech, occurrence in some dictionary. The task of a learning algorithm is to study the features over a large collection of annotated documents and identify rules that capture entities of a particular type.

3.1 Features

In our system, we have implemented the following groups of features:

Base-Line Features. This group includes features based mostly on the word's orthography: (1) word itself in lowercase; (2) word prefixes and suffixes of lengths 3-4; (3) word type: is-capitalized, all-capitalized, is-number, is-alphanumeric, contains-dash, contains-apostrophe, contains-digit, contains-dot, contains-capitalized-letter, is-punctuation-mark; (4) word parts before and after a dash in case of compound words; (5) whether the word is first in the sentence.

Morphological Features. These features are based on information provided by morphological disambiguator `t3mesta`: word lemma, POS-tag, word case, word ending, constituent morphemes.

Dictionary-based Features. We composed a large dictionary of entities covering common person names and surnames, local and international organizations and geographical locations. The dictionary contains entities in both Estonian and English. The lists of Estonian entities were obtained from multiple public on-line resources. A large collection of entities in English was downloaded from the web site of the Illinois Named Entity Tagger (Ratinov and Roth, 2009). Table 2 gives an overview of dictionary size and content. The dictionary covers 21% of the unique entities in the corpus, out of which 41% are unambiguous, meaning that the entity matches exactly one category in the dictionary.

Collected entities were preprocessed with a morphological disambiguator `t3mesta`. Words were replaced with their lemmas and turned to lower case. For a dictionary lookup we employed a leftmost longest match approach.

Dictionary Type	Size
Common Estonian first names (KeeleWeb, 2010)	5538
Common first and second names in English (Ratinov and Roth, 2009)	9348
Person full names in English (Ratinov and Roth, 2009)	877037
Estonian locations (Maa-amet, 2013)	7065
International locations in Estonian (Päll, 1999)	6864
Locations in English (Ratinov and Roth, 2009)	5940
Estonian organisations (Kaubandus-Tööstuskoda, 2010)	3417
International organisations (Ratinov and Roth, 2009)	329
Total	903279

Table 2: Dictionaries and numbers of entries.

WordNet Features. Estonian Wordnet is a knowledge base containing more than 27000 different concepts (sets of synonymous words) (Kerner et al., 2010). Wordnet encodes various semantic relationships between the concepts, which can be used as valuable information in NER tasks.

Based on the lemmas and their part-of-speech, we used Wordnet relations to encode hyperonymy, be in a state, belongs to a class and synset id information as extra features.

Global features. Global features enable to aggregate context from word's other occurrences in the same document (Chieu and Ng, 2003). We implemented global features as described in (Ratinov and Roth, 2009). For each occurrence $w_1, \ldots, w_N$ of the word w the set of features $c(w_i)$ is generated: (1) word is capitalized in document at any position, but the beginning of a sentence; (2) preceding word is a proper name; (3) following word is a proper name; (4) preceding word's presence in gazetteers; (5) following word's presence in gazetteers. Then, a set of features of the word w is extended with the aggregated context $\bigcup_{i=1}^{N} c(w_i)$.

3.2 Learning Algorithm

In this work, we use conditional random fields model (CRFs). CRFs are widely used for the task of NER due to their sequential nature and ability to handle a large number of features. Our choice is also substantiated by our earlier experiments on Estonian NER, where CRFs have demonstrated superior performance over a Maximum Entropy classifier (Tkachenko, 2010). We use CRFs implemented in the Mallet software package (McCallum, 2002).

4 Experiments and Results

In this section, we conduct a number of experiments to investigate the system behavior with respect to different factors.

We assess system performance using standard precision, recall and F_1 measure (Tjong Kim Sang and De Meulder, 2003). Scores for individual entity types are obtained by averaging results of 10-fold cross-validation on the full dataset. When splitting the data, document bounds are taken into account so that content of a single document fully falls either into training or test set. In this way, we minimize terminology transfer between samples used for training and testing. To summarize the results of an experiment with a single number, we report the weighted average of a corresponding measure over all entity types.

4.1 Named Entity Tag Representation

The choice of NE tag representation scheme has been shown to have significant effect on NER system performance (Ratinov and Roth, 2009). In this experiment, we set out to determine which scheme works best for the Estonian language. We consider two frequently used schemes – BIO (Ramshaw and Marcus, 1995) and BILOU. BIO format identifies each token as either the beginning, inside or outside of NE. BILOU format additionally distinguishes the last token of multi-token NEs as well as unit-length NEs. Hence, given NEs of three types (per, loc, org), the BIO scheme will produce 7 and BILOU 13 distinct tags.

Table 3 compares system performance using BIO and BILOU schemes. BILOU outperforms BIO in both precision and recall achieving a modest, but statistically significant 0.3 ppt improvement in F_1-score. This agrees with related research for the English language (Ratinov and Roth, 2009). In the following experiments we use a superior BILOU scheme.

Scheme	P (%)	R (%)	F_1 (%)
BIO	87.0	86.3	86.7
BILOU	**87.5**	**86.6**	**87.0**

Table 3: End system performance using BIO and BILOU tag representation schemes. BILOU outperforms BIO (p-value 0.04).

4.2 Feature Utility Analysis

	Feature group	P (%)	R (%)	F_1 (%)
1)	Baseline	83.3	76.8	79.9
2)	1) + Morphological	85.3	84.0	84.7
3)	2) + Dictionary	86.3	85.1	85.7
4)	2) + WordNet	85.4	84.2	84.8
5)	2) + Global	85.7	84.7	85.2
6)	All Features	**87.5**	**86.6**	**87.0**

Table 4: System performance using different groups of features.

Table 4 illustrates system performance using groups of features introduced in Section 3.1. We note that for each token we have included features from its immediate neighbors in the window of size 2. *Morphological* features demonstrate a major effect, increasing F_1-score by 4.8 ppt. Further inclusion of *Dictionary*, *WordNet* and *Global* features improves F_1-score by 1.0, 0.1 and 0.5 ppt respectively. By combining all groups of features, we achieve an overall F_1-score of 87%. Results for individual types of named entities are presented in Table 5. It is worth mentioning, that we have also attempted to do automatic feature selection using χ^2-test and by discarding infrequent features. However, both methods resulted in a significant loss of performance.

NE type	P (%)	R (%)	F_1 (%)
PER	**90.2**	**91.6**	**90.9**
ORG	80.0	74.7	77.1
LOC	89.4	89.6	89.5
ALL	87.5	86.6	87.0

Table 5: End-system performance.

4.3 Corpus Size

In this experiment, we study our system's learning capacity with respect to the amount of the training material. For this purpose, we repeat a 10-

fold cross-validation experiments with an increasing number of documents. In Figure 3, we observe the steepest gain in performance up to 300 documents, which further starts to flatten out. This indicates that our corpus is of an appropriate size for the task at hand, and that our system design is feasible.

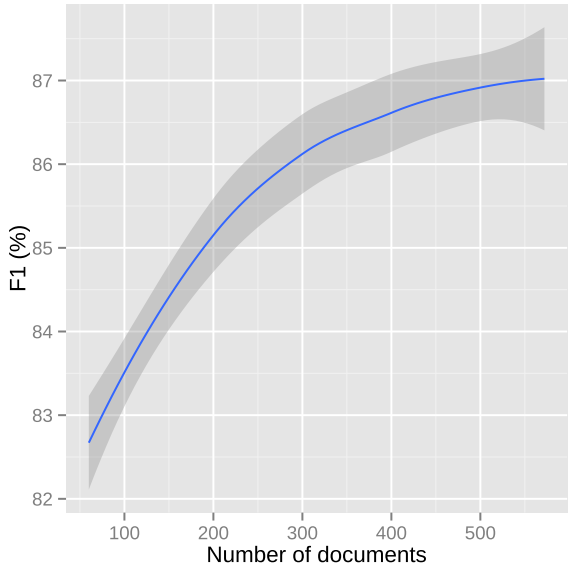

Figure 3: End-system smoothed F_1-score with increasing number of documents in the cross-validation corpus. Shaded area depicts 95% confidence interval.

4.4 NER without Morphological Analysis

In the previous section, we have shown that extending the baseline feature set with morphological features significantly boosts system performance. However, morphological analysis was performed with a commercial tool which may not be available due to licensing restrictions. It is, therefore, interesting to explore system performance without using such language specific features. In this experiment, we omit all the features produced by morphological analyzer. Since we still want to use *dictionary* and *global* features, we need to address an issue of word form normalization. For this purpose, we have built a simple statistical lemmatizer by analyzing lemmas and their inflected forms in Estonian Reference Corpus (Kaalep et al., 2010). As a result, we have achieved F_1-score of 84.8% – a 2.2 ppt decrease compared to the best result (see Table 6).

We conclude that even for highly inflective languages such as Estonian simple techniques for word form normalization, such as our lemmatizer, enable to achieve performance not much inferior than sophisticated linguistic tools.

lemmatizer	P (%)	R (%)	F_1 (%)
custom	86.4	83.3	84.8
t3mesta	**87.5**	**86.6**	**87.0**

Table 6: Performance comparison of NER systems using `t3mesta` and our custom-built lemmatizer.

5 Conclusions

In this work, we have addressed design challenges in building a robust NER system for Estonian. Our experiments indicate that a supervised learning approach using a rich set of features can effectively handle the complexity of the language. We demonstrated the importance of the features based on linguistic information, external knowledge and context aggregation. We observed that the choice of tag representation scheme affects system performance with BILOU outperforming a widely used BIO scheme. We also showed that an acceptable performance in NER can be achieved without using sophisticated language-specific linguistic tools, such as morphological analyzer. Last, but not least, we have built a first gold standard corpus for NER in Estonian and made it freely available for future studies. On this corpus, our system achieves an overall cross-validation F_1-score of 87%.

Acknowledgments

We would like to thank Filosoft[4] for kindly providing us morphological disambiguator `t3mesta`.

References

ACE. 2005. Automatic content extraction 2005 evaluation. Webpage: `http://www.itl.nist.gov/iad/mig//tests/ace/ace05/`.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 148–151. Association for Computational Linguistics.

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings*

[4] `http://www.filosoft.ee`

of the fifth conference on Applied natural language processing, pages 194–201. Association for Computational Linguistics.

Hai Leong Chieu and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. In *In Proceedings of the Seventh Conference on Natural Language Learning*, pages 160–163.

Nancy Chinchor. 1998. Muc-7 named entity task definition, version 3.5. In *Proc. of the Seventh Message Understanding Conference.*

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of COLING*, volume 96, pages 466–471.

Ralph Grishman. 1995. The NYU system for MUC-6 or where's the syntax? In *Proceedings of the 6th conference on Message understanding*, pages 167–175. Association for Computational Linguistics.

Heiki-Jaan Kaalep and Tarmo Vaino. 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. *Keel ja Kirjandus*, pages 30–38.

Heiki-Jaan Kaalep, Kadri Muischnek, Kristel Uiboaed, and Kaarel Veskis. 2010. The Estonian Reference Corpus: its composition and morphology-aware user interface. In *Proceedings of the 2010 conference on Human Language Technologies–The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 143–146. IOS Press.

Eesti Kaubandus-Tööstuskoda. 2010. List of Estonian organizations. Available at `http://www.koda.ee/?id=1916`.

KeeleWeb. 2010. List of common Estonian first names. Available at `http://www.keeleveeb.ee/`.

Kadri Kerner, Heili Orav, and Sirli Parm. 2010. Growth and revision of Estonian WordNet. *Principles, Construction and Application of Multilingual Wordnets*, pages 198–202.

Michal Konkol and Miloslav Konopík. 2011. Maximum entropy named entity recognition for Czech language. In *Text, Speech and Dialogue*, pages 203–210. Springer.

Dilek Küçük et al. 2009. Named entity recognition experiments on Turkish texts. In *Flexible Query Answering Systems*, pages 524–535. Springer.

Sven Laur, Alexander Tkachenko, and Timo Petmanson. 2013. Estonian NER corpus. Available at `http://metashare.ut.ee/repository/search/?q=Estonian+NER+corpus`.

Winston Lin, Roman Yangarber, and Ralph Grishman. 2003. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, volume 1, page 21.

Maa-amet. 2013. List of Estonian locations. Available at `http://www.maaamet.ee/index.php?lang_id=1&page_id=505`.

Elaine Marsh and Dennis Perzanowski. 1998. Muc-7 evaluation of IE technology: Overview of results. In *Proceedings of the seventh message understanding conference (MUC-7)*, volume 20.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 188–191. Association for Computational Linguistics. TEST.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Available at `http://mallet.cs.umass.edu/`.

Mārcis Pinnis. 2012. Latvian and Lithuanian named entity recognition with TildeNER. *Seed*, 40:37.

Peeter Päll. 1999. Maailma kohanimed. Eesti Keele Sihtasutus. Available at `http://www.eki.ee/knab/mkn_ind.htm`.

Lance A Ramshaw and Mitchell P Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge MA, USA.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 142–147. Association for Computational Linguistics.

Alexander Tkachenko. 2010. Named entity recognition for the Estonian language. Master's thesis, University of Tartu.

Dániel Varga and Eszter Simon. 2007. Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18(2):293–301.

Takahiro Wakao, Robert Gaizauskas, and Yorick Wilks. 1996. Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 418–423. Association for Computational Linguistics.