Segment Representations in Named Entity Recognition

Michal Konkol and Miloslav Konopík

Department of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia
Univerzitní 8, 306 14 Plzeň, Czech Republic
nlp.kiv.zcu.cz
{konkol,konopik}@kiv.zcu.cz

Abstract. In this paper we study the effects of various segment representations in the named entity recognition (NER) task. The segment representation is responsible for mapping multi-word entities into classes used in the chosen machine learning approach. Usually, the choice of a segment representation in the NER system is arbitrary without proper tests. Some authors presented comparisons of different segment representations such as BIO, BIEO, BILOU and usually compared only two segment representations. Our goal is to show, that the segment representation problem is more complex and that the proper selection of the best approach is not straightforward. We provide experiments with a wide set of segment representations. All the representations are tested using two popular machine learning algorithms: Conditional Random Fields and Maximum Entropy. Furthermore, the tests are done on four languages, namely English, Spanish, Dutch and Czech.

1 Introduction

Named entity recognition (NER) is a standard task of natural language processing. NER system searches for expressions of special meaning such as locations, persons, or organizations. These expressions often hold the key information for understanding the meaning of the document.

In this paper, we focus on one of many design aspects of a NER system: segment representation of multi-word entities. Many entities consist of multiple words (e.g. *Golan Heights*). If we use machine learning approach for NER, it is necessary to assign exactly one class to each token (word) in the corpus. The simplest way is to have one class for each type of named entity (and one extra type for normal words). This solution has a major limitation – it is not possible to correctly encode subsequent entities of the same type, e.g *"... the Golan Heights Israel captured from ..."* from CoNLL-2003 dataset where *Golan Heights* and *Israel* are both the *location* type. The result would look like this *"... word Location Location Location word word ..."*. Another motivation for more complex segment representations is that they can increase recognition performance (please note that we will use word *performance* in the meaning of an ability to recognize the named entities correctly not in the meaning of computational speed). For example, the recognition rules may differ for the first word and subsequent words of an entity. A segment representation that distinguishes the beginning of an entity then may

help with the recognition. The idea can be further extended by more complex segment representations.

In this paper we study effects of several segment representations on multiple languages using various machine learning (ML) approaches. The experiments with multiple languages are motivated by the fact that entities are very often proper names and different languages have very different rules for writing proper names. For example in English, all words except prepositions and conjunctions usually start with uppercase letter while in Czech only the first word starts with the uppercase letter, e.g. *Česká národní banka* in Czech and *Czech National Bank* in English. We use English, Spanish, Dutch and Czech corpora for our experiments.

Experiments with multiple ML approaches are important, because the optimal representation may vary for different methods. For our experiments we have chosen the Maximum Entropy (ME) [1] as a representative of classification methods and Conditional Random Fields (CRF) [2] as a representative of sequential methods.

The rest of the paper is organized as follows. Section 2 is devoted to the description of the segment representations. Section 3 is an overview of the related work. The NER system is described in Section 4. Section 5 gives a brief overview of the corpora used in our experiments. Section 6 describes our experiments and presents and discusses the results. The last section summarizes our findings.

2 Segment representations

As we have pointed out, there are multiple models for representing multi-word named entities (or more generally multi-word expressions). All the models (except the simplest one) use more than one tag for each type of named entity, e.g B-PERSON, I-PERSON for PERSON named entity. To our best knowledge, the most complex model uses 4 tags for each entity (plus one for not-an-entity tag). As already shown in the example, the tags are usually distinguished by a single letter prefix. The prefixes have a meaning of relative position in the named entity. The following list summarizes commonly used prefixes.

B (Beginning) Represents the first word of the entity.
I (Inside) Represents a part of the entity, which is not represented by other prefix.
L (Last, sometimes also **E**nd) Represents last word of the entity.
O (Outside or other) Represents word that is not a part of the entity.
U (Unit, sometimes also **W**ord or **S**ingle token) Represents a single word entities.

As we have said earlier, these models have two major purposes. The first one is to distinguish two subsequent entities. The model is able to do that, if it uses at least the **O**utside, **I**nside and one of the **B**egin and **E**nd tags. The second one, is to improve performance. Each tag is used as a single class in the ML methods. It means, that each tag represents a different set of statistics that can be used in the decision process. The intuition tells us, that the statistics accumulated over the corpus may be different for the first word of the entity (**B**), the inside word (**I**) and the other cases. For example the first word of the entity has much higher probability of having the first letter uppercase in Czech. In the following sections, we will describe the commonly used models.

IO model is the name we use for the simplest representation, even though this model has no well-known or widely accepted name. Each entity is represented only by one tag, which obviously does not need any prefix. This model is unable to decode subsequent entities of the same type, but it is not as important as it may seem at first sight, because subsequent entities of the same type are rare.

BIO model (or IOB) representation decodes each entity with two tags. There are two versions of the representation. The BIO-2 uses the **B**egin tag for each first word of an entity. The BIO-1 uses the **B**egin tag for the first word, only if it follows entity of the same type. In other words, the BIO-1 uses the **B**egin tag only if it has to distinguish subsequent entities.

IEO model is similar to the BIO representation, but it replaces the **B**egin tag with the **E**nd tag. There are also two versions – IEO-1 and IEO-2. These models have the same semantics as the BIO-1 and BIO-2 models.

BIEO model (BIOE, OBIE) representation uses both **B**egin and **E**nd tags.

BILOU model (C+O) representation is the most complex model used in NER. It adds the **U**nit tag for single word entities.

Furthermore, we experiment with newly created representations IOU, BIOU and OIEU models. They extend some of the previously mentioned models with the **U** tag.

3 Related work

The simplest segment representation (IO) was used by some of the first ML systems (e.g. [3–5]).

The CoNLL-2002 and CoNLL-2003 shared tasks used the BIO representation for annotations in their corpora (IOB-1 in 2002, IOB-2 in 2003) and many authors have adopted this model in their NER systems. The BIO model is the most commonly used model since these conferences.

The BIEO model was used in few papers [6–8], but it is very rare compared to the BIO model.

Some of the recent papers [9–11] adopted the BILOU representation probably based on the comparison in [10], where a comparison of the BIO and BILOU representation on English using CoNLL-2003 [12] and MUC-7 corpora using CRFs is provided . The BILOU representation performed better on the MUC-7 corpus on both (validation and test) data sets. On the CoNLL corpus, the BIO representation performed better on the validation set while the BILOU model performed better on the test set. The authors stated that segment representations can significantly impact the system and concluded that the BILOU model significantly outperforms the BIO model. The conclusion is only weakly supported by the results, in our opinion.

An interesting study is provided in [13]. The authors present method for using multiple segment representation together in one system. They also provide a comparison of multiple segment representations on the biomedical domain. The biomedical domain has different properties than the standard (news) corpora used in NER and cannot be compared with our results.

A similar research has been done for a different task – text chunking [14]. To our best knowledge, there are no other articles comparing segment representations in NER. We have not found any usage of representations not mentioned in this section.

4 NER system

We use two standard machine learning systems. The first one is based on Maximum Entropy (ME) and follows the description in [1]. The second one is based on Conditional Random Fields (CRF), similar to the baseline system in [15]. We use the Brainy ML library [16] for this purpose.

Both methods use the same feature set which consists of common NER features. The features are the following: words, bag of words, n-grams, orthographic features, orthographic patterns, and affixes.

5 Corpora

Our experiments are done on four languages – English, Spanish, Dutch and Czech. We use one corpus for each language.

For English, Spanish and Dutch we use the corpora from CoNLL-2002 and CoNLL-2003 shared tasks [17, 12]. These corpora have approximately 300,000 tokens and use four entity types – person (PER), organization (ORG), location (LOC) and miscellaneous (MISC).

For Czech we use the CoNLL format version of Czech Named Entity Corpus 1.1 [18, 19]. This corpus is smaller than the CoNLL corpora and has approximately 150,000 tokens. It uses 7 classes – time (T), geography (G), person (P), address (A), media (M), institution (I) and other (O).

All corpora use the BIO segment representation for the data. The English corpus (CoNLL-2003) uses the BIO-1 representation of segments. The rest the BIO-2. The segment representation of the corpora does not play any role in the training or evaluation as we firstly load the corpora to inner, corpus-independent representation and then transform it into training (or validation or test) data with proper segment representation for the given experiment.

6 Experiments

In all the experiments, we use the standard CoNLL evaluation with precision, recall and F-measure. We present only the F-measure because of space requirements. In the following sections we show two sets of experiments. The discussion of our results is in a separate section.

6.1 Standard partitioning

The first set of experiments is evaluated on the original partitioning of the corpora – training, validation and test set. For our experiments, we do not need to set any parameters based on the results on validation set. The results on the validation set thus provide the same information as on the test set. This follows the same procedure as in [10].

For each combination (segment representation, ml approach) we train a model on the training data and evaluate it on the validation and test data. The results of these experiments are shown in table 1.

Table 1: The results of our experiments on the standard partitioning of corpora.

(a) English

	ME		CRF	
	val	test	val	test
IO	86.89	78.66	88.98	83.64
IOU	**87.16**	**79.94**	88.75	83.60
BIO-1	86.89	78.27	88.98	83.61
BIO-2	86.86	79.15	88.08	83.74
BIOU	87.01	79.82	88.92	83.96
IEO-1	86.79	78.54	89.02	83.79
IEO-2	86.99	79.53	**89.25**	**84.16**
IEOU	86.91	79.85	89.10	83.62
BIEO	86.55	78.88	88.90	83.82
BILOU	86.42	79.50	88.84	83.47

(b) Spanish

	ME		CRF	
	val	test	val	test
IO	64.59	70.45	74.02	79.66
IOU	63.98	70.93	73.95	79.33
BIO-1	64.46	70.35	74.38	79.80
BIO-2	63.80	**71.37**	**74.56**	79.54
BIOU	64.04	71.27	74.15	79.18
IEO-1	**65.13**	71.03	74.27	**79.86**
IEO-2	63.12	70.33	74.45	79.50
IEOU	63.39	70.76	74.44	78.96
BIEO	63.30	70.54	74.46	79.55
BILOU	63.42	70.71	74.37	79.37

(c) Dutch

	ME		CRF	
	val	test	val	test
IO	67.25	70.09	74.31	76.34
IOU	69.28	71.85	74.39	76.62
BIO-1	67.49	70.06	**74.81**	76.53
BIO-2	68.31	70.45	74.37	76.23
BIOU	**69.56**	**72.53**	74.59	76.56
IEO-1	68.84	71.07	74.54	76.13
IEO-2	68.49	70.91	74.07	**77.17**
IEOU	69.23	72.34	73.63	76.79
BIEO	68.43	70.68	74.68	76.51
BILOU	68.78	72.08	73.82	76.82

(d) Czech

	ME		CRF	
	val	test	val	test
IO	56.93	53.48	**68.64**	68.41
IOU	57.26	54.33	68.12	68.05
BIO-1	56.16	53.45	68.50	68.90
BIO-2	56.96	54.99	68.44	69.11
BIOU	58.11	55.86	68.54	**70.26**
IEO-1	56.75	55.42	68.30	69.34
IEO-2	56.98	55.61	68.22	70.08
IEOU	58.21	56.64	67.92	69.55
BIEO	58.40	**57.21**	67.58	69.61
BILOU	**58.60**	56.73	67.41	69.21

6.2 Significance tests

The results of the first experiments are in many respects indecisive. For many representation pairs it is impossible to choose the better one (one is better on the test set, the other one on the validation set). Thus, we decided to perform a 10 fold cross-validation

to obtain more consistent results computed on much larger data. The advantage is that our tests do not depend on a short portion of data created by manual corpus division.

The data are prepared by the following procedure. Firstly, we concatenate all the data sets for each language (ordered: training, validation, test) into the data set D_{All} and number all the sentences (s denotes the index of a sentence). For fold i, $i = 0, \ldots, 9$, the test set is $D_{Test} = \{s : s \mod 10 = i\}$ and training set $D_{Train} = D_{All} - D_{Test}$. This procedure assures uniform distribution of sentences.

Each combination (segment representation, ML approach) is then tested on each fold. We compare the different combinations using the paired Student's t-test. The results are shown in Table 2 for ME and in Table 3 for CRF. We use two confidence levels $\alpha = 0.1, 0.05$. The null hypothesis H_0 is that there is no difference between segment representations. The alternative hypothesis H_1 is that one segment representation is significantly better than the other segment representation. Each cell contains four symbols, one for each language in the order English, Spanish, Dutch, and Czech.

- The symbol $<$ (resp. $>$) means, that the row segment representation is significantly worse (resp. better) than the column representation. The H_0 hypothesis is rejected at both levels $\alpha = 0.05, 0.1$.
- The symbol $\leq$ (resp. $\geq$) means, that the row representation is significantly worse (resp. better) than the column representation. The H_0 hypothesis is rejected at the level $\alpha = 0.1$, but we fail to reject it at the level $\alpha = 0.05$.
- The symbol $=$ is used for representations which are not significantly better or worse. We fail to reject hypothesis H_0.

Table 2: The significance tests for various segment representations using ME. Detailed description is provided in Section 6.2.

	IO	IOU	BIO-2	BIO-1	BIOU	IEO-2	IEO-1	IEOU	BIEO	BILOU
IO	====	====	>===	==≥=	==<<	==<=	<<=<	==≤≤	>===	=≥≤≤
IOU	====	====	>===	==>≥	==≤=	==<=	==>=	====	>===	>===
BIO-1	==≤=	==<≤	>=<=	====	==<<	==<=	=<=<	==<<	>===	=≥<<
BIO-2	<===	<===	====	<=>=	<=≤<	<≥<=	<=><	<==≤	>>==	=>=≤
BIOU	==>>	==≥=	>=≥>	==>>	====	===>	==>=	≤===	>=>≥	≥===
IEO-1	>>=>	==<=	>=<>	=>=>	==<=	=><>	====	==<=	>>=≥	≥><=
IEO-2	==>=	==>=	>≤>=	==>=	===<	====	=<><	=≤≥<	>=>=	>===
IEOU	==≥≥	====	>==≥	==>>	≥===	=≥≤>	==>=	====	>≥=≥	>>==
BIEO	<===	<===	<<==	<===	<=<≤	<=<=	<<=≤	<≤=≤	====	≤=<<
BILOU	=≤≥≥	<===	=<=≥	=≤>>	≤===	<===	≤<>=	<<==	≥=>>	====

6.3 Discussion

We start our discussion with the comparison to results of [10]. They compared BIO-1 and BILOU representations on the English CoNLL corpus using CRF. Our experiments have similar results. The BIO-1 representation was better on the test set, while

Table 3: The significance tests for various segment representations using CRF. Detailed description is provided in Section 6.2.

	IO	IOU	BIO-2	BIO-1	BIOU	IEO-2	IEO-1	IEOU	BIEO	BILOU
IO	====	>>=>	=<=>	≤<<≤	===>	=<<<	<==≤	==<=	=<=≥	>==≥
IOU	<<=<	====	<<=<	<<≤<	≤<≤<	<<<<	<==<	=<<<	=<=<	==<<
BIO-1	≥>>≥	>>≥>	>==>	====	>>=>	===≤	==>=	>>==	><>>	>==≥
BIO-2	=>=<	>>=>	====	<==<	>>==	==<<	<==<	≥>≤<	>===	>===
BIOU	===<	≥>≥>	<<==	<<=<	====	≤<<<	<==<	==<<	=<==	>===
IEO-1	>==≥	>==>	>==>	==<=	>==>	==<≤	====	>=<=	>==>	>==>
IEO-2	=>>>	>>>>	==>>	===≥	≥>>>	====	==>≥	≥>≥>	=<>>	>=>>
IEOU	==>=	=>>>	≤<≥>	<<==	==>>	≤<≤<	<=>=	====	=<>>	==>>
BIEO	=>=≤	=>=>	<===	<><<	=>==	=><<	<==<	=><<	====	=≥==
BILOU	<==≤	==>>	<===	<==≤	<===	<=<<	<==<	==<<	=≤==	====

the BILOU representation was better on the validation set. The differences slightly favor the BILOU representation, but it is unclear, if it is just a coincidence or the BILOU representation is better. This conclusion is also supported by the fact that in [10], the BILOU representation was better on the test set and worse on the validation set (in our case, better on the validation set, worse on the test set). The rest of the results has similar problems. For many representation pairs, it is impossible to pick the better one.

We were not satisfied with the results of the first set of tests, because it does not compare the representations rigorously. Thus, we proposed another approach for segment representations comparison described in Section 6.2. It is based on paired Student's test and gives well-defined comparisons.

The results of the significance tests are much more convincing. On one hand, the results provide evidence, that some segment representations are better than others. On the other hand, we are still unable to decide for many representations pairs, i.e. we must treat them as equal. Given these limitations, we can create a group of representations for each language, which are at the same or better level than all the other representations. These groups are (the bold representation has the highest average F-measure):

English, ME: IOU, BIO-1, IEO-1, **IEO-2**, IEOU
English, CRF: BIO-1, **IEO-1**, IEO-2
Spanish, ME: IOU, BIO-2, BIOU, **IOE-1**, IEOU
Spanish, CRF: BIO-2, IEO-1, **BIEO**
Dutch, ME: BIOU, **IEO-2**, BILOU
Dutch, CRF: BIO-1, **IEO-2**
Czech, ME: BIOU, **IEO-1**, IEOU, BILOU
Czech, CRF: IEO-2

Surprisingly, the IOE representations perform quite good, for Czech CRF is the IOE-2 even significantly better than the rest. The BILOU representation, generally considered as the best choice, performed rather poorly. We can say, that the optimal segment representation depends on both language and algorithm. We also expect it to be dependent on the feature set.

7 Conclusion

In this paper, we provide a rigorous study of segment representations for named entities. We experiment with ten different segment representations on the English, Spanish, Dutch and Czech corpora using two machine learning approaches – maximum entropy and conditional random fields.

We performed two sets of experiments. The first one was based on the standard partitioning of CoNLL corpora. The second one exploited 10 fold cross-validation and evaluation using the paired Student's t-test. The second test provides more accurate results.

Our experiments provide an interesting evidence. The BILOU representation ended up as the worst for English using CRF, even though it was considered better than the commonly used BIO-1 by [10] and it is generally considered as one of the best representations. The results presented in [10] were similar to the results of our first set of experiments, but the second set of experiments disproved this hypothesis. The IOE-1 and IOE-2 representations seem to be the best or at least reasonable choice for almost all languages and methods. Surprisingly, these representations have not been used in NER yet.

We show that choosing the optimal segment representation for named entities is a complex problem. The optimal representation depends on the language (corpus), on the approach, and very likely on the feature set. We propose a well-defined procedure for finding the optimal representation.

Thus, the impact of the article is two fold. First, we propose a new procedure for segment representation evaluation. Second, we recommend the use of IOE-1 and IOE-2 as they provide the most promising results in our tests.

In the future, we would like to experiment with multiple feature sets and their relation to optimal segment representation. The relation of the data size and the optimal representation could be also interesting.

References

1. Borthwick, A.E.: A Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York, NY, USA (1999) AAI9945252.
2. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
3. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Proceedings of the fifth conference on Applied natural language processing. ANLC '97, Stroudsburg, PA, USA, Association for Computational Linguistics (1997) 194–201
4. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999) 100–110
5. Béchet, F., Nasr, A., Genet, F.: Tagging unknown proper names using decision trees. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. ACL '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 77–84

6. Cucerzan, S., Yarowsky, D.: Language independent ner using a unified model of internal and contextual evidence. In: Proceedings of, Taipei, Taiwan (2002) 171–174
7. Mao, X., Xu, W., Dong, Y., He, S., Wang, H.: Using Non-Local Features to Improve Named Entity Recognition Recall. Volume 21., The Korean Society for Language and Information (KSLI) (2007)
8. Sun, J., Wang, T., Li, L., Wu, X.: Person name disambiguation based on topic model. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing. (2010)
9. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing named entities in tweets. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 359–367
10. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. CoNLL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 147–155
11. Straková, J., Straka, M., Hajič, J.: A new state-of-the-art czech named entity recognizer. In Habernal, I., Matouek, V., eds.: Text, Speech, and Dialogue. Volume 8082 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 68–75
12. Tjong Kim Sang, E.F., De Meulder, F.: 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 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 142–147
13. Cho, H.C., Okazaki, N., Miwa, M., Tsujii, J.: Named entity recognition with multiple segment representations. Information Processing & Management **49**(4) (2013) 954 – 965
14. Shen, H., Sarkar, A.: Voting between multiple data representations for text chunking. In Kgl, B., Lapalme, G., eds.: Advances in Artificial Intelligence. Volume 3501 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2005) 389–400
15. Lin, D., Wu, X.: Phrase clustering for discriminative learning. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. ACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1030–1038
16. Konkol, M.: Brainy: A machine learning library. In Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M., eds.: Artificial Intelligence and Soft Computing. Volume 8468 of Lecture Notes in Computer Science. Springer International Publishing (2014) 490–499
17. Tjong Kim Sang, E.F.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. COLING-02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 1–4
18. Konkol, M., Konopík, M.: Crf-based czech named entity recognizer and consolidation of czech ner research. In Habernal, I., Matoušek, V., eds.: Text, Speech and Dialogue. Volume 8082 of Lecture Notes in Computer Science., Springer Berlin Heidelberg (2013) 153–160
19. Ševčíková, M., Žabokrtsky, Z., Krůza, O.: Named entities in czech: annotating data and developing ne tagger. In: Proceedings of the 10th international conference on Text, speech and dialogue. TSD'07, Berlin, Heidelberg, Springer-Verlag (2007) 188–195