Hybrid chunking for Turkish combining morphological and semantic features

Razieh Ehsani, Ercan Solak, and Olcay Taner Yıldız

Işık University, Istanbul, Turkey

Abstract. We use morphological features together with the semantic representations of words to solve chunking problem in Turkish. We separately train and tune word embeddings for semantic representations and conditional random fields for morphological features. We combine the two in a random forest.

1 Introduction

Chunking is considered as an early form of parsing and defined as dividing a sentence into syntactically disjoint sets of meaningful word-groups or chunks [1]. An example of a sentence split up into chunks is shown below:

(1) [By midday] [NP the London market] [VP was in full retreat]

In Example (1), the chunks are represented as groups of words between square brackets, and the tags denote the type of the chunk.

Extracted chunks can be used as input in more complex natural language processing tasks, such as information retrieval, document summarization, question answering, statistical machine translation, etc. Compared to syntactic and/or dependency parsing, chunking is an easier problem. For this reason, in the earlier years of statistical natural language processing, many researchers put emphasis on the chunking problem [2].

In manually tagging a corpus, human annotators intuitively combine their linguistic knowledge with their competence of the language. In our approach, we try to capture these two sources of expertise and native knowledge about a language. We train a CRF with morphological features to learn the linguistic knowledge. For the competence, we train a neural network to represent the words as numeric vectors in an high dimensional space.

The paper is organized as follows: In Section 2, we give related work on chunking in Turkish and other languages. We describe our approach in Section 3. We report the experimental results in Section 4 and discuss them in Section 5. We conclude in Section 6.

2 Related Work

Ramshaw and Mitchell (1995) work [3] was one of the earliest works on chunking. They applied transformation-based learning approach for NP chunks on the data derived from Penn-Treebank. [4] applied support vector machines (SVM) to identify NP chunks.

Using an ensemble of 8 SVM-based systems, they got 93% in terms of F-measure on CoNLL shared task data. The work in other languages, especially on less resourced and/or morphologically rich languages, is scarce. There are limited works on Korean [5, 6], Hindi [7], and Chinese [8, 9].

Kutlu's works [10] and [11] proposed the first Turkish NP chunker based on a dependency parser with handcrafted rules. NPs are divided into two sub-classes as main NPs (base NPs) and all NPs (including sub-NPs). Kahlout work [12] annotated verb chunks in METU-Sabancı Turkish dependency treebank [13] and replaced phrase chunks with constituent chunks. The remaining chunks are left as general chunks and only their boundaries are detected. The algorithm is based on conditional random fields (CRF) and achieves a best accuracy of 91.95 for verb chunks and 87.50 for general chunks.

3 Combining nominal features with word embeddings

For the modeling part, we train the semantic and morphological models separately and combine them in a random forest. Each part is tuned to maximize its performance in the chunking task when used in isolation.

3.1 Semantic representation

We restrict ourselves to the semantics of a word as defined by the contexts in which it is used. To encode the context information we embed the words in real vectors using `word2vec` algorithm described in [14]. In order to customize the embeddings towards our task, we tuned the free parameters, window size and the embedding dimension of `word2vec` that maximize the performance in a partial evaluation of chunking.

3.2 Morphological features

In order to train a learner for the morphological features in isolation, we use a CRF with a set of nominal features that are derived from a word. CRFs are known to perform well in sequential labeling tasks with nominal contextual features, [15].

For languages like Turkish where the syntactic functions of the words are often indicated by the case suffixes, morphological features become crucial for automatic labeling tasks. A particular example that is common in chunking is the genitive-possessive NP structure in Turkish. The genitive case marker and the possessive morpheme often also mark the boundaries of a NP. Even when the genitive possessive NP is composed of other NPs, anything between genitive marker and the possessive morpheme is often inside the outer NP.

Another set of case markers that are important in identifying propositional phrases in Turkish are the dative, locative and ablative markers, -A, -DA and -DAn. In English, starting boundaries of PPs are often marked with function words like "to," "at" and "from" These function words roughly correspond to the markers for the three noun cases just mentioned. As Turkish is a head final language, these three cases indicate the right boundary of a PP.

Many verbs in Turkish are formed by combining a noun with one of the two auxiliary verbs et- (do) and ol- (be). Examples are “yardım etmek” (to help), “kabul etmek” (to accept), “neden olmak” (to cause), “ait olmak” (to belong). For such cases, the auxiliary verb and noun preceding it must be chunked together. For surface forms we define the ETOL binary feature which indicates whether the root is et- or ol-.

The full set of features used in the isolated morphological chunking is given in Table 1. The contribution of the addition of each feature to the performance is given in the experiments in Table 1.

Table 1: Morphological features

Identifier	Definition
F0	Previous case
F1	Current surface form
F2	Previous surface form
F3	Next is ETOL, binary
F4	Previous POS
F5	Current is possessive, binary
F6	Previous is possessive, binary
F7	Current case
F8	Current is ETOL, binary

3.3 Combining semantics and morphology

Among the nominal features that we used in the CRF chunking were surface forms of the word itself and its surrounding words. Considering the effective vocabulary size of a morphologically rich language like Turkish, the surface forms in a training set make a rather sparse set of features. The same is to lesser extent true for the root forms of the words.

On the other hand, the morphological features have values from a small set, often binary.

Instead of surface forms, we used their embedded representations as real vectors. This brings into the learner the semantics contained in the word embeddings. Besides, using vectors alleviate the problems of sparsity.

CRFs as commonly used for sequential labeling do not handle continuous valued features without a pre-processing such as binning. In order to directly combine the nominal features with continuous vector features, we use random forest.

4 Experiments

For training word2vec, we used the Milliyet corpus of Turkish which contains more than 5M sentences. For training and testing in the chunking task, we used the translated

sentences of a portion of Penn Treebank with about 8K sentences [16]. The first 90% is used for training and the rest for testing. We generated the chunk labels automatically using the root label of the largest subtree under the tree root.

We experimented with two sets of chunk labels. The simple set B, I only identifies the boundaries of the chunks. The larger set identifies the class of the chunks as well.

In order to find the most suitable window and vector sizes in `word2vec` for chunking task, we set up an SVM classifier that uses only the vector representation of word at t as the feature in deciding the chunk label at t. For the SVM and random forest implementations, we used the one provided by Weka, [17].

We used the simple label set $\{$B, I$\}$. The average F-scores are shown in Table 2. Comparing the F-scores, we chose a window size of 1 and an embedding size of 20. We used these parameters for the rest of the experiments.

Table 2: The average F-scores for different embedding and window size combinations.

vector	window		
	1	2	3
5	68.1	66.2	65.7
10	69.8	69.3	67.9
20	**70.3**	69.3	68.7
50	69.4	69.4	68.2

Next, we determined the relevant context window for the surface forms of the words around the current word w_t. The largest window is $\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$. We tested the random forest classifier with concatenated embeddings for window sizes of 1, 2 and 3 within this larger context. The weighted F-scores as percentages are given in Table 3. We see that the most relevant surface forms for chunking are $\{w_{t-1}, w_t, w_{t+1}\}$, one word to each side of the current word to be tagged.

Table 3: The average F-scores for different contexts around the word to be tagged.

w_{t-2}	w_{t-1}	w_t	w_{t+1}	w_{t+2}	F-score
		✓	✓	✓	73
	✓	✓	✓		**84.6**
✓	✓	✓			81.8
		✓	✓		74
	✓	✓			82.7
		✓			71.70

To find the best morphological features for chunking, we used a CRF where for each set of candidate features, we also included the surface forms $\{w_{t-1}, w_t, w_{t+1}\}$. For a fast implementation of CRF, we used `wapiti`, [18]. Table 4 shows the incremental contribution of each feature to the overall performance. We used the floating feature selection algorithm [19] to determine the set of features in Table 1.

Table 4: Feature selection for morphological chunking.

Feature set	Description	F-score
S0	F0	74.06
S1	S0 $\cup \{F1\}$	78.84
S2	S1 $\cup \{F2\}$	80.31
S3	S2 $\cup \{F3\}$	81.93
S4	S3 $\cup \{F4\}$	83.16
S5	S4 $\cup \{F5\}$	83.66
S6	S5 $\cup \{F6\}$	83.89
S7	S6 $\cup \{F7\}$	84.19
S8	S7 $\cup \{F8\}$	84.42

Finally, we combine the semantics and morphological features by using the best morphological feature set of CRF by replacing the surface forms with their embeddings. The resulting set contains both numerical and nominal features. We used a random forest to combine these. The performance for the set of simple labels is given in Table 5.

Table 5: Performance of the combined features for the set of simple chunking labels, B and I.

Label	F-score
B	84
I	88.9

The performance and the confusion matrix for the set of full labels are given in Table 6 and Table 7. In order to save space, we show a partial confusion matrix that contains only the most salient errors.

Table 6: Performance of the combined features for the full set of chunking labels.

Label	F-score
B-CC	0.882
B-VG	0.738
B-NP	0.684
I-NP	0.632
B-ADVP	0.578
I-ADVP	0.491
I-PP	0.446
I-VG	0.411
I-S	0.395
B-PP	0.317
B-S	0.168
I-ADJP	0.095
B-ADJP	0
I-CC	0

Table 7: Confusion matrix for the full set of labels.

	classified as								
	B-NP	I-NP	B-VG	B-PP	I-PP	I-ADVP	I-VG	B-S	I-S
B-NP	390	46	13	15	3	0	2	10	19
I-NP	42	648	8	9	56	3	20	3	52
B-VG	25	21	282	6	2	0	12	2	16
B-PP	51	24	6	42	9	1	1	5	11
I-PP	10	**129**	1	6	127	4	3	0	31
I-ADVP	4	11	1	3	4	27	3	0	2
I-VG	5	72	41	3	14	4	77	0	19
B-S	56	32	3	9	3	0	0	15	10
I-S	28	**185**	27	14	32	10	10	5	162

5 Discussion

We see that the highest performance is for CC which identifies conjunctions. This is somewhat expected as there are only a few conjunctions in Turkish and they can easily be identified through their surface forms.

The next best performance is for the start of verb groups (VG). Most verb groups in Turkish are just single words. The case of the previous word is an indicative feature in distinguishing the start of a verb group.

In English parsing, VP groups together the object NP and several PP and ADVP phrases under the same VP tree. However, for Turkish, such a grouping is too coarse. In generating our training and test data out of translated PennTreebank sentences, we automatically identified and extracted the verb under VP and chunked it as a verb group, VG. Then, we identified the remaining subtrees of VP tree as chunks with their own labels.

The confusion matrix in Table 7 highlights the distribution of errors in different chunks. Most salient error source is the wrong classification of words within a NP as either being inside a PP or S. PP errors must be due to the insufficiency of case marking features to differentiate between stand alone NPs and NPs within PPs. Similarly, many sub-clauses in Turkish is bounded by genitive cases marking which is similar to genitive possessive NP constructions.

6 Conclusion

In this work, we proposed a new hybrid approach to Turkish chunking. We combined morphological features of the words together their semantic representations as real valued vector embeddings. We trained a CRF and a random forrest separately to determine the best feature set and the best surface context around a word to be tagged. We combined the selected nominal features and surface embeddings in a random forest to achieve a hybrid chunker.

For training CRF, we intentionally refrained from using the root forms of the words as it is not obvious how we would replace them with their vector embeddings because word embeddings were estimated from a corpus of surface forms. As a future study, we plan to device a method to meaningfully embed the roots and thus enrich the hybrid feature set.

References

1. Abney, S.: Parsing by chunks. In: Principle-Based Parsing, Kluwer Academic Publishers (1991) 257–278
2. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence). 2 edn. Prentice Hall (2009)
3. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Third ACL Workshop on Very Large Corpora. (1995) 82–94
4. Kudo, T., Matsumoto, Y.: Chunking with support vector machines. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. NAACL '01, Stroudsburg, PA, USA, Association for Computational Linguistics (2001) 1–8
5. Park, S.B., Zhang, B.T.: Text chunking by combining hand-crafted rules and memory-based learning. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. ACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 497–504
6. Lee, Y.H., Kim, M.Y., Lee, J.H.: Chunking using conditional random fields in korean texts. In: Proceedings of the Second International Joint Conference on Natural Language Processing. IJCNLP'05, Berlin, Heidelberg, Springer-Verlag (2005) 155–164

7. Gune, H., Bapat, M., Khapra, M.M., Bhattacharyya, P.: Verbs are where all the action lies: Experiences of shallow parsing of a morphologically rich language. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 347–355
8. Chen, W., Zhang, Y., Isahara, H.: An empirical study of chinese chunking. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions. COLING-ACL '06, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 97–104
9. Sun, G.L., Huang, C.N., Wang, X.L., Xu, Z.M.: Chinese chunking based on maximum entropy markov models. International Journal of Computational Linguistics & Chinese Language Processing **11** (2006) 115–136
10. Kutlu, M.: Noun phrase chunker for Turkish using dependency parser. Master's thesis, Sabancı University (2010)
11. Kutlu, M., Cicekli, I.: Noun phrase chunking for turkish using a dependency parser. In: Lecture Notes in Electrical Engineering. ISCIS' 2015, Springer Verlag (2014) 381–391
12. El-Kahlout, I.D., Akin, A.A.: Turkish constituent chunking with morphological and contextual features. In: CICLing (1). (2013) 270–281
13. Atalay, N.B., Oflazer, K., Say, B.: The annotation process in the Turkish treebank. In: 4th International Workshop on Linguistically Interpreted Corpora. (2003)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR. (2013)
15. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 134–141
16. Yıldız, O.T., Solak, E., Görgün, O., Ehsani, R., AS, A.L.T.T.: Constructing a turkish-english parallel treebank. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 2. (2014) 112–117
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11** (2009) 10–18
18. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (2010) 504–513
19. Alpaydın, E.: Introduction to Machine Learning. 3^{rd} edn. The MIT Press (2015)