English Nominal Compound Detection with Wikipedia-Based Methods*

István Nagy T.[1], Veronika Vincze[2]

[1] University of Szeged, Department of Informatics,
6720 Szeged, Árpád tér 2., Hungary
[2] MTA-SZTE Research Group on Artificial Intelligence,
6720 Szeged, Tisza Lajos krt. 103., Hungary
{nistvan,vinczev}@inf.u-szeged.hu

Abstract. Nominal compounds (NCs) are lexical units that consist of two or more elements that exist on their own, function as a noun and have a special added meaning. Here, we present the results of our experiments on how the growth of Wikipedia added to the performance of our dictionary labeling methods to detecting NCs. We also investigated how the size of an automatically generated silver standard corpus can affect the performance of our machine learning-based method. The results we obtained demonstrate that the bigger the dataset, the better the performance will be.

Keywords: Wikipedia, multiword expressions, nominal compounds, MWE detection, silver standard corpus

1 Introduction

In natural language processing, multiword expressions (MWEs) have been receiving special interest. Nominal compounds (NCs) form a subtype of multiword expressions: they form one unit the parts of which are meaningful units on their own, the unit functions as a noun and it usually has some extra meaning component compared with the meanings of the original parts [1]. The semantic relation between the parts of the nominal compound may vary: it may express a "made of" relation (*apple juice*), a "location" relation (*neck pain*) or a "made for" relation (*hand cream*) just to name a few. Thus, nominal compounds encode some important meaning components that can be fruitfully applied by e.g. information extraction systems. However, such applications like these require that nominal compounds should be previously known to the system.

Nominal compounds occur frequently in everyday English (in the Wiki50 corpus [2], 67.3% of the sentences on average contain a nominal compound). Furthermore, they are productive: new nominal compounds are entering the language all the time, hence they cannot be exhaustively listed and appropriate methods should be implemented for their identification.

It is also important to emphasize that a nominal compound candidate does not always function as a nominal compound. Take, for instance, *tall boy*: when it refers to

* This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

a can of beer, it is an MWE, but when it refers to a young male of somewhat unusual height, it is simply a productive combination of an adjective and a noun and does not constitute an MWE. Thus, nominal compounds should be identified in context, i.e. in running texts, and we will follow this approach in our investigations.

2 Related Work

The identification of MWEs or more specifically nominal compounds has been received considerable attention. Bonin et al. [3] use contrastive filtering in extracting multiword terminology (mostly nominal compounds) from scientific, Wikipedia and legal texts: term candidates are ranked according to their belonging to the general language or the sub-language of the domain. Caseli et al. [4] use alignment-based techniques to extract multiword expressions from parallel corpora in the pediatrics domain. Nagy T. et al. [5] describe a rule-based method, which heavily relies on morphological information to identify nominal compounds in Wikipedia texts. The machine learning-based tool `mwetoolkit` is designed to extract MWEs from texts, which is illustrated by extracting English nominal compounds from the Genia and Europarl corpora and from general texts [6, 7]. In this paper, we present our experiments to automatically detect English nominal compounds in running texts with Wikipedia-based machine learning methods similar to [8] and investigate how the extension of Wikipedia contributes to the process.

3 Experiments

For the evaluation of our models, we made use of two corpora. First, we used Wiki50 [2], in which several types of multiword expressions (including nominal compounds) and Named Entities were marked. This corpus consists of 50 Wikipedia pages, and contains 2929 occurrences of nominal compounds. We also investigated approaches on the 1000-sentence dataset from the British National Corpus that contains 485 two-part nominal compounds [9]. The dataset includes texts from various domains such as literary work, essays, newspaper articles etc. Statistical data on the corpora can be seen in Table 1.

Corpus	**Sentences**	**Tokens**	**Nominal Compounds**	**2**	**3**	**4≤**
Wiki50	4,350	114,570	2929	2442	386	101
BNC dataset	1,000	21,631	485	436	40	9

Table 1. Corpora used for evaluation with the number of tokens of the nominal compounds, based on their length.

3.1 Wikipedia-based Method for Detecting Nominal Compounds

To identify nominal compounds we used a Wikipedia-based approach similar to Vincze et al. [2]. They collected lowercase n-grams from English Wikipedia links, and automatically filtered the non-English terms, Named Entities and non-nominal compounds.

They combined three methods in the following way: a candidate was marked as a nominal compound if it occurred in the list of n-grams. The second method involved the merging of two possible nominal compounds; namely if `A B` and `B C` both occurred in the list, `A B C` was also accepted as a nominal compound. Third, a nominal compound candidate was marked if it occurred in the list and its Part of Speech (POS)-tag sequence matched one of the previously defined patterns (e.g. `adjective + noun`). POS-tags were determined by the Stanford POS Tagger [10]. Finally, they combined these three methods, and this combined approach proved to be the most successful. This is why we applied this method later on.

Vincze et al. [2] investigated the performance of their Wikipedia-based method only on an actual Wikipedia state. However, we thought it interesting to examine this approach from the beginning of Wikipedia and to investigate how the size of Wikipedia influences the results. Hence, we collected the above mentioned nominal compound list from the actual Wikipedia state of the beginning of each year. The English Wikipedia was launched in 2001, so the first list was collected from the state of 1 January 2002.

3.2 Machine Learning Approaches

In order to automatically identify nominal compounds, we also applied a machine learning-based method [2]. The tool uses the MALLET implementations [11] of the Conditional Random Fields (CRF) classifier [12]. Identifying multiword Named Entities and nominal compounds can be carried out in a similar way as both nominal compounds and multiword Named Entities consist of more than one words. They form one semantic unit and thus, they should be treated as one unit in NLP systems [8]. Therefore the feature set employed was developed on the basis of a general Named Entity feature set, which includes the following categories: **orthographical features:** capitalization, word length, bit information about the word form (contains a digit or not, has an uppercase character inside the word, etc.), character level bi/trigrams, suffixes; **dictionaries** of first names, company types, denominators of locations; **frequency information:** frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token, which was derived from the Gigaword dataset; **shallow linguistic information:** part of speech; **contextual information:** sentence position, trigger words (the most frequent and unambiguous tokens in a window around the word) from the training database and the word between quotes.

This basic feature set was extended with features adapted to nominal compounds. The **dictionaries** were extended with different nominal compound lists. We collected a nominal compound list from the state of Wikipedia on 1 January 2013 and sorted it according to frequency of occurrence. The components with different frequencies were included in different dictionaries. In addition, the training and test sets of Task 9 of the SemEval 2010 [13] were used as dictionaries. The shallow linguistic features were extended with the **POS-rules**, so if the POS-tag sequence in the text matched one pattern typical of nominal compounds (e.g. `noun – plural noun`), the sequence tags were marked as *true*, otherwise *false*. Furthermore, the **other entities** were also specified in the sentence, like Named Entities (NEs) or Light Verb Constructions (LVCs),

which were also used as features. To identify Named Entities, the Stanford Named Entity Recognition (NER) tool was applied [14] and we detected Light Verb Constructions similar to the method described in [5].

We trained the first-order linear chain CRF classifier with the above mentioned feature set and evaluated it on the two corpora in a 10-fold cross-validation setting at the sentence level. We trained the CRF models with the default settings in Mallet for 200 iterations or until convergence was reached. We applied the above mentioned dictionary-based method to automatically generate a silver standard corpus. In this case, the training set consisted of randomly selected Wikipedia pages, which do not contain lists, tables or other structured texts. These documents were not manually annotated, so the dictionary-based nominal compound labeling was treated as the silver standard. The resulting dataset is much bigger than the available manually annotated corpora, but the annotation is less reliable. In this case, we would like to exploit the fact of the big training data with less accurate annotation.

The CRF model was trained on the silver standard dataset with the above presented feature set. We investigated how the size of the automatically labeled training set influenced the performance of CRF. First, we analyzed the results when the training set only consisted of 10 Wikipedia pages. After, we gradually increased the automatically labeled training set with randomly selected Wikipedia pages.

As we used randomly selected Wikipedia pages to train our CRF model, we investigated how the random selection affected the results. We automatically generated ten different training sets. One set consisted of ten thousand randomly selected Wikipedia pages, where dictionary-based labeling was used as the silver standard and a CRF model was trained with the above described feature set.

We also compared the results achieved by the supervised leave-one-document-out model, the model trained on the automatically generated dataset, and the dictionary-based method on the Wiki50 corpus.

4 Results

Table 2 shows the results obtained by the dictionary-based approach, the number of Wikipedia pages and the size of the collected lists, depending on the years and the actual state of Wikipedia.

After the first year, the English Wikipedia only consisted of 13,200 pages, and we were able to extract 5,892 potential nominal compounds from the links and the dictionary-based method, which yielded an F-score of 9.52 on the Wiki50 corpus. At the beginning of 2013, the English Wikipedia consisted of 9,914,544 pages, the potential NC list contains 687,574 elements and the approach achieved an F-score of 56.59. As Table 2 shows, with the expansion of Wikipedia, the method managed to produce better results, but the rate of improvement is negligible after 2007. Moreover, in 2013 the dictionary-based method yielded an F-score that was 0.15 lower than that in 2012.

We also investigated how the training set size affected the results of the model trained on the automatically generated dataset. As Figure 1 shows, with an increased training set the machine learning approach could achieve better results, but the improvement was smaller. The method produced an F-score of 46.69 when the training set

Year	WikiPages	NC list	Recall	Precision	F-score	Diff.
2002	13,200	5,892	5.12	68.42	9.52	-
2003	124,229	25,431	16.22	59.05	25.45	+15.93
2004	271,160	58,696	24.99	71.69	37.06	+11.61
2005	752,239	120,028	33.81	69.57	45.50	+8.44
2006	1,611,876	211,802	40.11	66.20	49.96	+4.46
2007	2,988,703	322,918	44.42	64.15	52.49	+2.53
2008	4,432,034	405,635	46.91	63.35	53.90	+1.41
2009	5,281,708	459,544	48.51	62.82	54.74	+0.84
2010	6,009,776	511,303	49.33	62.45	55.12	+0.38
2011	7,167,621	567,288	50.69	62.66	56.04	+0.92
2012	9,007,810	640,879	53.36	60.58	56.74	+0.7
2013	9,914,544	687,574	53.67	59.84	56.59	-0.15

Table 2. The results of applying the Wikipedia-based dictionary labeling method, depending on the expansion of Wikipedia in terms of recall, precision, and F-score. **WikiPages:** the number of Wikipedia pages. **NC list:** the size of the lists collected from the Wikipedia links.

just consisted of 10 Wikipedia pages and an F-score of 56.06 when it was constructed from 10,000 Wikipedia pages.

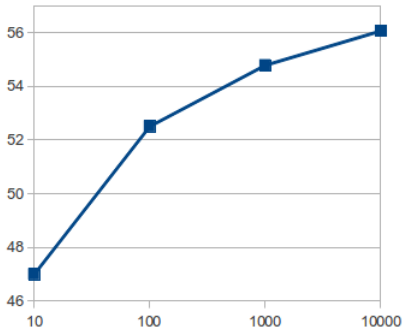

Fig. 1. Results of the machine learning approach depending on the automatically generated training set size (the number of Wikipedia pages).

Table 3 lists ten different CRF model results, trained on ten different automatically generated datasets. The average F-score of ten runs was 55.99 and the standard deviation was 0.3237. Table 4 gives the results of the different approaches for the Wiki50 and BNC datasets.

To perform an error analysis, we examined the length of nominal compounds in the corpora. As Table 1 shows, the Wiki50 corpus contains 83.37% (2442 occurrences) two-part, 13.17% (386 occurrences) three-part nominal compounds and only 3.46%

	Recall	Precision	F-score
1	57.02	55.21	56.1
2	56.74	55.38	56.05
3	57.26	55.73	56.48
4	56.64	55.02	55.82
5	57.46	55.25	56.33
6	56.88	55.61	56.24
7	56.98	55.03	55.99
8	56.2	54.94	55.56
9	57.08	53.73	55.36
10	56.85	55.04	55.93
avg.:	56.91	55.1	55.99

Table 3. Machine learning results obtained on different automatically generated training sets in terms of recall, precision, and F-score in Wiki50

(101 occurrences) are longer than three tokens. As for the BNC dataset, there are 436 (89.89%) two-part nominal compounds, and only 8.25% (40 occurrences) are three-part, while 1.86% (9 occurrences) contain more than three tokens. Table 4 shows that all the methods got their best results on the two-part nominal compounds. Longer nominal compounds yielded worse results in the case of each method and corpus.

	LOO Wiki50	WikiTrain Wiki50	Dict. Wiki50	Wikitrain BNC	Dict. BNC
2	69.12/79.62/74.00	64.86/60.14/62.41	61.14/64.66/62.85	40.60/45.04/42.70	33.49/45.06/38.42
3	52.33/62.93/57.14	29.02/47.86/36.13	30.05/49.79/37.48	20.00/22.86/21.33	17.50/17.95/17.72
4≤	24.73/45.10/31.94	8.60/40.00/14.16	6.45/75.00/11.88	0.00/0.00/0.00	0.00/0.00/0.00
All	64.39/72.40/68.16	56.57/55.57/56.06	53.67/59.84/56.59	38.02/41.53/39.70	31.40/40.75/35.47

Table 4. Results of different methods for nominal compounds in terms of recall, precision, and F-score in Wiki50 corpus and BNC dataset. **LOO:** evaluated in the leave-one-document-out scheme. **WikiTrain:** CRF model trained on the automatically generated dataset. **Dict:** Wikipedia-based dictionary labeling.

Table 4 also reveals that on the Wiki50 corpus the CRF model evaluated with the leave-one-document-out scheme yielded the best results with an F-score of 68.16. The CRF model trained on the automatically generated dataset and the Wikipedia-based dictionary labeling method achieved the same F-score on the Wiki50 corpus with different recall and precision scores. The machine learning-based method yielded a higher recall with a lower precision. Moreover, this approach yielded an F-score that was 4.23 higher on the BNC dataset than the dictionary labeling method.

5 Discussion

Due to the dynamic expansion of Wikipedia, the dictionary-based method was able to extract bigger potential nominal compound lists from Wikipedia links and achieved better recall scores for each year. At the same time, while the automatically extracted

list was noisy, the precision score continuously decreased over the years, but the F-score value increased up to 2013. Then in 2013 the rise in recall was less than the decrease in precision, hence the F-score value was lower for 2013 than that for 2012. As Table 2 shows after 2009 the F-score improvement was less than 1. However, we found that the dynamic expansion of Wikipedia had a positive effect on the recall score, so in order to improve the precision score, we should define stronger rules.

Next, we evaluated the machine learning-based model with the leave-one-document-out scheme on the Wiki50 corpus. This approach achieved the highest F-score value since we used a supervised model here. As we applied a silver standard corpus, we had a less accurate but much bigger training dataset where the automatic (therefore noisy) labeling was used as a silver standard. This method had a detrimental effect on the precision scores for the CRF model, but recall scores improved because the model had access to more labeled nominal compounds. We examined how the training set size influenced the performance of this machine learning-based approach and we found that the size of training data had a large impact on the performance of the method when we exploited the automatically generated training data. We also wanted to see how the random selection of Wikipedia pages affected the performance. The method proved to be sufficiently robust as the standard variation of F-score values was 0.3237.

We also examined the nature of English nominal compounds, and we found that the majority of nominal compounds are two-part and the investigated approaches performed well on the two-part compounds as opposed to longer compounds, which is probably due to the fact that automatically labeled examples contained fewer instances of longer compounds.

On the BNC dataset the machine learning method proved to be more effective than the dictionary-based method. Due to the BNC paper [9] they annotated sequences of two nouns. However, we found 40 three-part, and 9 longer nominal compounds too in the data. On the other hand, some of the errors are related to annotation errors, for instance, marking nominal compounds that contain a proper noun, e.g. *Belfast primary school headmaster*, as simple nominal compounds instead of proper nouns (as they should be according to the guidelines). These differences can be responsible for the weaker performance of our methods on the BNC dataset.

6 Conclusions

Here, we examined dictionary and machine learning-based methods for identifying nominal compounds in two corpora. These approaches made intensive use of Wikipedia data. The dictionary-based approach applied a list automatically collected from Wikipedia. We examined the results of this method that depended on the expansion of Wikipedia over the years. We found that the growth of Wikipedia improved the performance, especially the recall score, but the rate of improvement is decreased over time. We also looked at the effectiveness of the machine learning-based method when it was trained on an automatically generated silver standard corpus and we demonstrated that this approach can also provide acceptable results. In the future, we would like to improve the precision of automatic labeling as this will have a positive effect on the performance of both the machine learning approach and the dictionary labeling method.

References

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico (2002) 1–15
2. Vincze, V., Nagy T., I., Berend, G.: Multiword expressions and named entities in the Wiki50 corpus. In: Proceedings of RANLP 2011, Hissar, Bulgaria (2011)
3. Bonin, F., Dell'Orletta, F., Venturi, G., Montemagni, S.: Contrastive filtering of domain-specific multi-word terms from different types of corpora. In: Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, China, Coling 2010 Organizing Committee (August 2010) 77–80
4. Caseli, H.d.M., Villavicencio, A., Machado, A., Finatto, M.J.: Statistically-driven alignment-based multiword expression identification for technical domains. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore, ACL (August 2009) 1–8
5. Nagy T., I., Vincze, V., Berend, G.: Domain-dependent identification of multiword expressions. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee (September 2011) 622–627
6. Ramisch, C., Villavicencio, A., Boitet, C.: mwetoolkit: a framework for multiword expression identification. In: Proceedings of LREC'10, Valletta, Malta, ELRA (May 2010)
7. Ramisch, C., Villavicencio, A., Boitet, C.: Web-based and combined language models: a case study on noun compound identification. In: Coling 2010: Posters, Beijing, China (August 2010) 1041–1049
8. Nagy T., I., Berend, G., Vincze, V.: Noun compound and named entity recognition and their usability in keyphrase extraction. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee (September 2011) 162–169
9. Nicholson, J., Baldwin, T.: Interpreting Compound Nominalisations. In: LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco (2008) 43–45
10. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of EMNLP 2000, Stroudsburg, PA, USA, ACL (2000) 63–70
11. McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
12. 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
13. Erk, K., Strapparava, C., eds.: Proceedings of the 5th International Workshop on Semantic Evaluation. ACL, Uppsala, Sweden (July 2010)
14. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 363–370