Classification of Noun-Noun Compound Semantics in Dutch and Afrikaans

Ben Verhoeven, Walter Daelemans
CLiPS – Computational Linguistics Group
University of Antwerp
Antwerp, Belgium
{Ben.Verhoeven;Walter.Daelemans}@ua.ac.be

Gerhard B van Huyssteen
CTexT – Centre for Text Technology
North-West University
Potchefstroom, South Africa
Gerhard.Vanhuyssteen@nwu.ac.za

***Abstract*—This article presents initial results on a supervised machine learning approach to determine the semantics of noun compounds in Dutch and Afrikaans. After a discussion of previous research on the topic, we present our annotation methods used to provide a training set of compounds with the appropriate semantic class. The support vector machine method used for this classification experiment utilizes a distributional lexical semantics representation of the compound's constituents to make its classification decision. The collection of words that occur in the near context of the constituent are considered an implicit representation of the semantics of this constituent. F-scores were reached of 47.8% for Dutch and 51.1% for Afrikaans.**

Keywords—compound semantics; Afrikaans; Dutch; machine learning; distributional methods

I. INTRODUCTION

Computational language understanding can be seen as one of the major goals of research in computational linguistics and natural language processing (NLP). However, many issues need to be resolved before we can even approximate human level language understanding. A notable obstacle, for example, is the productivity that a language exhibits in creating new words. An important and very productive word formation process, in at least Germanic languages, is compounding [1:141]. Since these new words are not available in a computational dictionary and their meanings are hence not explicated, a computational system will have trouble interpreting the meaning of these words. Existing NLP applications, such as question answering, information extraction and machine translation systems, will benefit from better compound understanding. This paper presents initial results on first-generation semantic analyzers for Dutch and Afrikaans noun-noun compounds.

This research builds to a great extent on techniques previously used and discussed by Ó Séaghdha [2] for English and Verhoeven [3] for Dutch. Some results of the latter are revisited in this article.

The structure of this paper will be as follows. First, a summary of related research on the topic will be presented. This summary will focus on the techniques used in our own research. We then describe our annotation scheme and process for the Dutch and Afrikaans noun-noun compounds. The classification experiments are then discussed, after which we present our results and propose some directions for further research.

II. RELATED RESEARCH

Past research on semantic analysis of noun-noun compounds has focused almost exclusively on English. The problem of semantically analyzing these compounds was mostly considered a supervised machine learning problem. Different approaches were proposed considering two main characteristics of the research: the scheme of categories being used for the semantic classification of the compounds, and the features that the machine learning algorithm uses to classify the compounds.

A. Classification Schemes

Several attempts have been made in the past to come up with appropriate classification schemes for noun-noun compound semantics. These schemes are mainly inventory-based in that they present a limited list of predefined possible classes of semantic relations a compound can have. Early work in computational research is due to Warren [4], Finin [5] and Lauer [6].

In some cases, proposed classes are abstractly represented by a paraphrasing preposition as in [6], [7] and [8]. For example, all compounds that can be paraphrased by putting the preposition 'of' between the constituents belong to the class OF, e.g. a 'car door' is the 'door of a car'. Another possibility is using predicate-based classes where the relations between the constituents are not merely described by a preposition but by definitions or paraphrasing predicates for each class. The class AGENT would contain compounds that could be paraphrased as 'X is performed by Y' [9], e.g. *enemy activity* can be paraphrased as 'activity is performed by the enemy'. Different schemes vary from 9 to 43 classes with kappa scores for inter-annotator agreement ranging from 52% to 62% [2][4][7] [10][11][12][13][14].

B. Features

With regard to the information used by the classifier to assign the classes to the compounds, two main roads are

available, *viz.* taxonomy-based methods, or corpus-based methods.

Taxonomy-based methods (also called semantic network similarity [15]) base their features on a word's location in a taxonomy or hierarchy of terms. Most of the taxonomy-based techniques use WordNet [16] for these purposes; especially the hyponym information in the hierarchy is used. A bag of words is created of all hyponyms and the instance vector contains binary values for each feature (the feature being whether the considered word from the bag of words is a hyponym of the constituent or not). Kim and Baldwin reached an accuracy of 53.3% using only WordNet [9]. Other research was based on Wikipedia as a semantic network [17] or the MeSH hierarchy of medical terms [18].

Corpus-based methods use co-occurrence information of the constituents of the selected compounds in a corpus. The underlying idea – the distributional hypothesis – is that the set of contexts in which a word occurs, is an implicit representation of the semantics of this word [17]. This information can be used in different ways. Ó Séaghdha [2] describes measures of lexical similarity and relational similarity.

The lexical similarity measure assumes that compounds are semantically similar when their respective constituents are semantically similar. The co-occurrences of both constituents will be combined to calculate a measure of similarity for the entire compound. Accuracies[1] of 54.98% [12][17] and 61% have been reached [2][20].

The relational similarity measure assumes two pairs of constituents "to be similar if the contexts in which the members of one pair co-occur are similar to the contexts in which the members of the other pair co-occur" [2:118]. Ó Séaghdha and Copestake [17] report an initial accuracy of 42.34%. This result was improved to 52.6% in [2]. Lapata and Keller [8] report an accuracy of 55.71% with web-based relational similarity. Their corpus-based similarity's accuracy was only 27.85%.

Nastase *et al.* [21] extract grammatical collocations of the constituents from a corpus and use it as features for the classifier. This collocation includes words that appear with the target word in a grammatical relation, e.g. subject, object, etc.

Corpus-based and taxonomy-based methods have also been combined by several researchers. Accuracies of 58.35% [19], 79.3% [12] and even 82.47% [21] were reported.

III. Annotation

In order to perform a supervised machine learning experiment, we need semantic information of compounds that machine learning algorithms can learn from. There is thus a need for examples with an explicit description of the compound semantics, as is created through manually annotating data.

The compounds considered for manual annotation are only those noun-noun compounds that do not occur in a dictionary – otherwise a semantic classification is both unnecessary and unwanted: unnecessary because there is already a gloss for the compound present (the meaning is thus already known), and unwanted because we want to train our classifier on the systematics that will be found in the semantics of newly produced compounds. However, the constituents of these compounds are required to appear in a dictionary. If the constituents would not be present in a dictionary, their individual meanings would not be known to us and semantically relating an unknown word to some other word seems pointless. Hence, compounds with proper nouns (e.g. *Beneluxland* 'Benelux country') will be excluded from our dataset.

A. Scheme and Guidelines

For our research, we adopted the annotation scheme and guidelines created by Ó Séaghdha [2], which were by and large based on Levi's set of categories from 1987 [2]. The guidelines were developed for semantic annotation of English noun-noun compounds, so some adaptations were in order. These adaptations mainly existed of supplementing the guidelines with Dutch and Afrikaans examples. More details on other changes can be found in [3].

The annotation tag of each compound consists of three parts: the category, the annotation rule by which the category is determined, and the direction in which the rule applies. The annotation scheme will be summarized here; the complete guidelines can be found on the project website[2].

Ó Séaghdha [2] describes eleven classes of compounds; six of these classes are semantically specific. These classes include:

- BE: The compound can be rewritten as 'N2 which is (like) (a) N1' with N1 and N2 being the two constituents nouns. Example: *woman doctor*
- HAVE: The compound denotes some sort of possession. Part-whole compounds, typical one-to-many possession, compounds expressing conditions or properties and meronymic compounds belong here. Example: *car door*
- IN: The compound denotes a location in time or place. Example: *garden party*
- ACTOR: The compound denotes a characteristic event or situation and one of the constituents is a salient entity. Example: *enemy activity*
- INST: The compound denotes a characteristic event and there is no salient entity present. Example: *cheese knife*
- ABOUT: The compound describes a topical relation between its constituents. Example: *film character*

1 The accuracies presented in the related research section are mentioned as an indication of those systems' performance. Comparison with our own results is not in order due to the use of different data, methods, etc.

2 http://tinyurl.com/aucopro

The other five categories are less specific. The MISTAG and NONCOMPOUND categories serve to classify compounds that do not belong in the dataset. MISTAG refers to the fact that one or both of the constituents is not a common noun (e.g. *London Town*, where N1 is a proper noun). NONCOMPOUNDs are not two-noun compounds (e.g. 'a salt and *pepper beard*'). The REL class describes compounds with a clear meaning that does not belong to any of the other classes, but of which the relation between the constituents seems productive (e.g. *sodium chloride*). The LEX category is almost the same as REL, but the relation does not seem to be productive (e.g. *monkey business*). The UNKNOWN category is for correct noun-noun compounds of which the meaning is not clear enough to annotate.

B. Dutch

The Dutch noun-noun compounds were taken from a compound list that was extracted from the e-Lex Dutch lexicon[3]. This compound list was already split into constituents and the POS tags of the constituents were available. The WNT (Woordenlijst Nederlandse Taal) lexicon [22] was used to check the occurrence of the compounds and constituents in a dictionary. The eventual compound list contained 1802 Dutch noun-noun compounds.

The Dutch compound set was annotated by a student in linguistics that played no role in the development of the annotation guidelines. One of the authors of this paper annotated a subset of 500 compounds to be able to calculate an inter-annotator agreement (IAA). Both annotators are native speakers of Dutch. The reported IAA was 60.2% (Kappa = 0.60) [3].

C. Afrikaans

The Afrikaans noun-noun compounds were taken from the CKarma list of splitted compounds [23]. Since there were no POS tags available, these compounds were manually selected from the list. These compounds and their constituents were not crosschecked with a dictionary; this will be the case in future research. The compound list contained 1500 Afrikaans noun-noun compounds.

The complete Afrikaans compound set was annotated by three bachelor students in language, all native speakers of Afrikaans. The pair-wise average IAA was 53.4% (Kappa = 0.53). This IAA is a bit lower than our IAA for Dutch, possibly due to the fact that lexicalized compounds were not removed from the annotation list. They might be harder to annotate because their lexicalized meaning is not always a logical semantic relation between their constituents and may not fit into one of our categories then. Take the Afrikaans *naaldenkoker* as example; this compound has 'needle case' as literal meaning, but it also has a lexicalized meaning: 'dragonfly'. It is clear that lexicalized compounds may cause annotation difficulties.

3 This compound list was created by Lieve Macken of the LT3 research group at University College Ghent.

IV. Experiment

The conducted experiments were based on those conducted by Ó Séaghdha [2]. We will provide a description of our own experimental setup here. An in-depth discussion of the methodology and more extensive experimentation on the Dutch data can be found in [3].

Our classification experiment is based on a combination of the distributional hypothesis (as proposed above) with the idea of analogical reasoning. It is assumed that the semantic category of a compound can be predicted by comparing compounds with similar meanings [2].

A. Lexical Similarity

The lexical similarity measure is a corpus-based method of feature selection. As described above, this measure will compare the semantic similarities of the constituents of the considered compounds. The modifiers of the compounds (normally the left-hand members of the compound) will be compared with each other and the compound heads (normally the right-hand members of the compound) will be compared with each other. Two compounds, for example 'flour can ' and 'corn bag' will be considered similar if they have similar modifying constituents ('flour' and 'corn') and similar head constituents ('can' and 'bag'). In this example, the similarity would be rather high because the compounds both denote a container with its content.

B. Vector Creation

In order to perform a classification experiment, one needs the information for each instance (in this case: each compound) to be stored in a vector. This section will describe the creation of these vectors.

1) Bag of words (BOW)

For every compound constituent, the co-occurrence context was calculated. For this purpose, for each instance of the constituents in the corpus, the surrounding *n* words (that belong to the 10,000 most frequent words of the corpus) were held in memory. The number of context words was 3 or 5 to both the left and right hand side of the constituent in the two variants of the experiment. The relative frequencies of these context words (the number of times the word appeared in the context of the constituent, divided by the frequency of the constituent in the corpus) for each constituent were stored.

For Dutch, the Twente News Corpus [24] was used. This is a 340 million word corpus of newspaper articles. For Afrikaans, we used the Taalkommissie corpus [25], a 60 million word corpus that consists of a variety of text genres.

A concatenation of the constituent data is used to create the instance vector features. Each instance vector contains the compound it represents, its category, direction and annotation rule, and the relative frequencies for the 1000 most frequent words for each constituent (hence 2000 per compound). However, for purposes of training data in our experiment, the vectors are stripped from their compound, direction and rule, leaving only the category and the features. Compounds of which one or both of the constituents did not appear in the corpus were excluded from the data.

The classification experiment dealt with those compounds that are annotated with a semantically specific category. This means that only compounds with the category tags BE, HAVE, IN, INST, ACTOR and ABOUT were used for the experiments. The final vector set for Afrikaans contains 1439 compounds, while the final vector set for Dutch has 1447 compounds. The class distributions for Dutch and Afrikaans are presented in Table 1.

TABLE I. CLASS DISTRIBUTIONS FOR DUTCH AND AFRIKAANS

	Dutch		**Afrikaans**	
BE	105	7.3%	359	25.0%
HAVE	233	16.1%	140	9.7%
IN	428	29.5%	299	20.8%
ACTOR	62	4.3%	126	8.8%
INST	235	16.2%	108	7.5%
ABOUT	384	26.6%	407	28.2%
Total	1447		1439	

2) Principal Component Analysis

The BOW approach that was described so far takes the occurrence of each word as one attribute in the vector. Our vectors thus have 2000 attributes and one class (the category) each. This makes our experimentation computationally rather expensive. Principal component analysis (PCA) was used to reduce the dimensionality of our vectors to improve the performance of our system.

Performing PCA on a matrix or vector of data transforms this data by mathematically optimizing the variance between the instances. The vectors will reduce in size because correlated attributes will be fused into new attributes that are called principal components (PCs) [3:42].

The 'PCA Module for Python', as implemented by Risvik [26] was used to perform these mathematical transformations on our data. Apart from our BOW vectors, we now also have a PCA vector for both context variants.

C. Machine Learning

For the actual machine learning experiments on the four sets of vectors (BOW and PCA, each with 3 or 5 context words), we used the SMO algorithm, which is WEKA's [27] support vector machines (SVM) implementation. Automatic optimization of the parameters was performed by the CVParameterSelection function.

We used 10-fold cross-validation; the classifier was trained and tested ten times on a different train and test set. The ten folds cover the whole data set maximally. The average results and standard variation of these ten runs are a representation of the performance of this classifier.

V. RESULTS

Since this is the first research on both Dutch and Afrikaans, we will assume the most frequent class probability in the datasets as baselines for these classifiers. This baseline is calculated by dividing the count of the most frequent class by the total number of compounds in the dataset. This number represents the accuracy that can be obtained by always guessing this most frequent class as the output class. For Dutch, this baseline is 29.5% (428 instances of class IN on a total of 1447 compounds) [3]. For Afrikaans, this baseline is 28.2% (407 instances of class ABOUT on a total of 1439 instances).

TABLE II. RESULTS OF SMO CLASSIFIER ON DUTCH COMPOUND SEMANTICS

	Precision	**Recall**	**F-Score**
BOW 3	47.6	48.0	47.8
PCA 3	41.7	46.2	41.7
BOW 5	47.7	48.0	47.8
PCA 5	43.0	47.6	43.6

All results in Table 2 of the classification experiment with Dutch compounds show a significant improvement over the most frequent class baseline (29.5%). The BOW approach seems to do better than the PCA results with an F-score of 47.8% for both the 3 and 5 word variant. The results for the PCA approach (41.7% and 43.6%) are somewhat lower, but still significantly higher than the baseline.

TABLE III. RESULTS OF SMO CLASSIFIER ON AFRIKAANS COMPOUND SEMANTICS

	Precision	**Recall**	**F-Score**
BOW 3	50.8	51.6	51.1
PCA 3	47.7	50.5	47.5
BOW 5	50.3	50.8	50.5
PCA 5	49.3	51.3	48.5

Table 3 shows that the classification experiment with Afrikaans compounds also performs significantly better than its most frequent class baseline of 28.2%. The highest F-score reached was 51.1% for the BOW approach with 3 context words. These results are even slightly better than our results for Dutch.

This 3% improvement of the Afrikaans over the Dutch performance may be ascribed to the final annotation list for Afrikaans being a combination of the semantic annotations of three persons. In taking the most agreed upon class for each compound, we may have reached a better approximation of the actual compound semantics than when using the annotation list of just one person, as we did for Dutch. However, this hypothesis remains a subject for further research.

VI. CONCLUSION AND FURTHER WORK

This paper presented, for the first time, exploratory research on the semantic classification of noun-noun compounds in Dutch and Afrikaans. The results show that a first approach, based on corpus-based semantic representations, already provides promising results for both Afrikaans (highest F-score of 51.1%) and Dutch (highest F-score of 47.8%). Although a full comparison with earlier systems for English is not appropriate, we can note that the results of our initial classifiers already compare favorably to previous results for English; for example, Ó Séaghdha reaching an F-score of 58.8% (accuracy

of 61%) also using only lexical similarity with a training set of 1443 compounds [2].

The performance of the classifiers significantly outperforms the most frequent class baselines. The BOW approach turns out to provide better results than the PCA approach, because it seems that some of the information in the vectors is lost during PCA calculation. It is nevertheless our intention to further explore the PCA approach and variants in future research, because the computational performance of the approach is important in practical applications. We will also investigate alternative methods for constructing corpus-based lexical semantic representations, explore the use of lexical databases (a lexical semantic network such as WordNet is also available for Dutch, while a small-scale WordNet of Afrikaans is also available), and experiment with context-based representations.

We will try and test other machine learning algorithms, such as memory-based learning. An attempt will be made to improve the IAA's as well.

The semantics of other compounds than noun-noun compounds, such as verb-noun and adjective-noun compounds, will be investigated from a linguistic perspective, in order to determine the viability to model such semantic relations computationally.

ACKNOWLEDGMENT

The current paper fits in a broader research on automatic compound processing. Automatic Compound Processing (AuCoPro) is a mutual project by research groups of the North-West University (Potchefstroom, South Africa), the University of Antwerp (Belgium) and Tilburg University (The Netherlands). The University of Antwerp deals mainly with the compound semantics subproject, Tilburg University deals mainly with compound splitting. North-West University works on the Afrikaans aspects of both subprojects.

This research was co-funded by a joint research grant of the Nederlandse Taalunie (Dutch Language Union) and the Department of Arts and Culture (DAC) of South Africa and a grant of the National Research Foundation (NRF) (grant number 81794).

We also want to acknowledge the work of the bachelor students of the North-West University, Potchefstroom Campus (Carli de Wet, Nadia Schultz, Benito Trollip and Joanie Liversage) that annotated the Afrikaans compounds as part of their bachelor dissertation.

REFERENCES

[1] G. Booij, The Morphology of Dutch, Oxford: Oxford University Press, 2002.

[2] D. Ó Séaghdha, "Learning compound noun semantics," Ph.D. thesis, University of Cambridge, UK, 2008.

[3] B. Verhoeven, "A computational semantic analysis of noun compounds in Dutch," M.A. thesis, University of Antwerp, Belgium, 2012.

[4] B. Rosario, and M. Hearst, "Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy," in Proc. EMNLP, 2001, 82-90.

[5] T. W. Finin, "The semantic interpretation of compound nominal," in Proc. AAAI, 1980.

[6] M. Lauer, "Designing statistical language learners," Ph.D. thesis, Macquarie University, Australia, 1995.

[7] R. Girju, D. Moldovan, M. Tatu, and D. Antohe, "On the semantics of noun compounds," in Computer Speech and Language, vol. 19, 2005, pp.479-496.

[8] M. Lapata, and F. Keller, "The web as a baseline: evaluating the performance of unsupervised web-based models for a range of NLP tasks," in Proc. NAACL-HLT, 2004, pp. 121-128.

[9] S. N. Kim, and T. Baldwin, "Automatic interpretation of noun compounds usin WordNet similarity," in Proc. IJCNLP, 2005, pp. 945-956.

[10] P. Nakov, "Noun compound interpretation using paraphrasing verbs: feasability study," in Proc. AIMSA, 2008.

[11] D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju, "Models for the semantic classification of noun compounds," in Proc. NAACL-HLT Workshop on Computational Lexical Semantics, 2004, pp. 60-67.

[12] S. Tratz, and E. Hovy, "A taxonomy, dataset, and classifier for automatic noun compound interpretation," in Proc. ACL, 2010, pp. 678-687.

[13] K. Barker, and S. Szpakowicz, "Semi-automatic recognition of noun-modifier relationships," in Proc. ICCL, 1998, pp. 96-102.

[14] D. T. Wijaya, and P. Gianfortoni, "'Nut-case: what does it mean?': Understanding semantic relationship between nouns in noun compounds through paraphrasing and ranking the paraphrases," in Proc. SMER, 2011.

[15] D. Ó Séaghdha, "Semantic classification with WordNet kernels," in Proc. NAACL-HLT Short Papers, 2009, pp. 237-240.

[16] G. A. Miller, "WordNet: a lexical database for English," Communication of the ACM, vol. 38, 1995, pp. 39-41.

[17] D. Ó Séaghdha, and A. Copestake, "Co-occurrence contexts for noun compound interpretation," in Proc. Workshop on a Broader Perspective on Multiword Expressions, 2007, pp. 57-64.

[18] Z. Harris, Mathematical Structures of Language. New York: Interscience, 1968.

[19] D. Ó Séaghdha, "Annotating and learning compound noun semantics," in Proc. ACL Student Research Workshop, 2007, pp.73-78.

[20] D. Ó Séaghdha, and A. Copestake, "Semantic classification with distributional kernels," in Proc. COLING, 2008, pp. 649- 656.

[21] V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz, "Learning noun-modifier semantic relation with corpus-based and WordNet-based features," in Proc. AAAI, 2006, pp. 781-787.

[22] Nederlandse Taalunie, "Bronbestand woordenlijst Nederlandse taal," Internet: http://www.inl.nl/tst-centrale/nl/producten, 2005 [18/09/2012].

[23] CTexT, CKARMA ("C5 KompositumAnaliseerder vir Robuuste Morfologiese Analise") [C5 Compound Analyser for Robust Morphological Analysis]. Potchefstroom: Centre for Text Technology (CTexT), North-West University, 2005.

[24] R. Ordelman, F. de Jong, A. Van Hessen, and H. Hondorp, "TwNC: a multifaceted Dutch news corpus," ELRA Newsletter, vol. 12, pp. 3-4, 2007.

[25] Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns, Taalkommissiekorpus 1.1. Potchefstroom: Centre for Text Technology (CTexT), North-West University, 2011.

[26] H. Risvik, "PCA module for Python," Internet: http://folk.uio.no/henninri/pca_module, 2008 [27/05/2012].

[27] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (Third Edition). Burlington, MA: Morgan Kaufmann, 2011.