"Could you *make* me a favour and *do* coffee, please?": Implications for Automatic Error Correction in English and Dutch

Sophia Katrenko
UiL-OTS
Utrecht University
s.katrenko@uu.nl

Abstract

The correct choice of words has proven challenging for learners of a second language and errors of this kind form a separate category in error typology. This paper focuses on one known example of two verbs that are often confused by non-native speakers of Germanic languages, *to make* and *to do*. We conduct experiments using syntactic information and immediate context for Dutch and English. Our results show that the methods exploiting syntactic information and distributional similarity yield the best results.

1 Introduction

When learning a second language, non-native speakers make errors at all levels of linguistic analysis, from pronunciation and intonation to language use. Word choice errors form a substantial part of all errors made by learners and may also be observed in writing or speech of native speakers. This category of errors includes homophones. Some commonly known confusions in English are *accept-except*, *advice-advise*, *buy-by-bye*, *ate-eight*, to name but a few. Other errors can be explained by a non-native speaker's inability to distinguish between words because there exists only one corresponding word in their native language. For example, Portuguese and Spanish speakers have difficulties to differentiate between *te doen (to do)* and *te maken (to make)*, and Turkish between *kunnen (can)*, *weten (to know)* and *kennen (to know)* in Dutch (Coenen et al., 1979). Adopting terminology from Golding and Roth (1999) and Rozovskaya and Roth (2010), *do/make* and *kunnen/kennen/weten* form two confusion sets. However, unlike the case of *kunnen/kennen/weten*, where the correct choice is often determined by syntactic context [1], the choice between *to make* and *to do* can be motivated by semantic factors. It has been argued in the literature that the correct use of these verbs depends on what is being expressed: *to do* is used to refer to daily routines and activities, while *to make* is used to describe constructing or creating something. Since word choice errors have different nature, we hypothesize that there may exist no uniform approach to correct them.

State-of-the-art spell-checkers are able to detect spelling and agreement errors but fail to find words used incorrectly, e.g. to distinguish *to make* from *to do*. Motivated by the implications that the correct prediction of two verbs of interest may have for automatic error correction, we model the problem of choosing the correct verb in a similar vein to selectional preferences. The latter has been considered for a variety of applications, e. g. semantic role labeling (Zapirain et al., 2009). Words such as *be* or *do* have been often excluded from consideration because they are highly polysemous and "do not select strongly for their arguments" (McCarthy and Carroll, 2003). In this paper, we study whether semantic classes of arguments may be used to determine the correct predicate (e.g., *to make* or *to do*) and consider the following research questions:

1. Can information on semantic classes of direct

[1] *Kunnen* is a modal verb followed by the main verb, *kennen* takes a direct object as in, e.g., *to know somebody*, and *weten* is often followed by a clause (as in *I know that*).

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 49–53,
Montréal, Canada, June 7-8, 2012.

objects potentially help to correct verb choice errors?

2. How do approaches using contextual and syntactic information compare when predicting *to make* vs. *to do*?

The paper is organised as follows. Section 2.1 discusses the methods, followed by Section 2.2 on data. The experimental findings are presented in Section 2.3. We conclude in Section 3.

2 Experiments

We re-examine several approaches to selectional preferences in the context of error correction. Existing methods fall into one of two categories, either those relying on information from WordNet (McCarthy and Carroll, 2003), or data-driven (Erk, 2007; Schulte im Walde, 2010; Pado et al., 2007). For the purpose of our study, we focus on the latter.

2.1 Methods

For each verb in question, we have a frequency-based ranking list of nouns co-occurring with it (verb-object pairs) which we use for the first two methods.

Latent semantic clustering (LSC) Rooth et al. (1999) have proposed a soft-clustering method to determine selectional preferences, which models the joint distribution of nouns n and verbs v by conditioning them on a hidden class c. The probability of a pair (v, n) then equals

$$P(v,n) = \sum_{c \in C} P(c)P(v|c)P(n|c) \qquad (1)$$

Similarity-based method The next classifier we use combines similarity between nouns with ranking information and is a modification of the method described in (Pado et al., 2007). First, for all words n_i on the ranking list their frequency scores are normalised between 0 and 1, f_i. Then, they are weighed by the similarity score between a new noun n_j and a corresponding word on the ranking list, n_i, and the noun with the highest score (1-nearest neighbour) is selected:

$$\arg\max_{n_i} f_i \times sim(n_j, n_i) \qquad (2)$$

Finally, two highest scores for each verb's ranking list are compared and the verb with higher score is selected as a preferred one.

In addition, if we sum over all seen words instead of choosing the nearest neighbour, this will lead to the original approach by Pado et al. (2007). In the experimental part we consider both approaches (the original method is referred to as **SMP** while the nearest neighbour approach is marked by **SMknn**) and study whether there is any difference between the two when a verb that allows many different arguments is considered (e.g., it may be better to use the nearest neighbour approach for *to do* rather than aggregating over all similarity scores).

Bag-of-words (BoW) approach This widely used approach to document classification considers contextual words and their frequencies to represent documents (Zellig, 1954). We restrict the length of the context around two verbs (within a window of ± 2 and ± 3 around the focus word, *make* or *do*) and build a Naive Bayes classifier.

2.2 Data

Both verbs, *to make* and *to do*, license complements of various kinds, e. g. they can be mono-transitive, ditransitive, and complex transitive (sentences 1, 2, and 3, respectively). Furthermore, *make* can be part of idiomatic ditransitives (e.g., *make use of*, *make fun of*, *make room for*) and phrasal mono-transitives (e.g., *make up*) .

1. Andrew made [a cake]$_{dobj}$.
2. Andrew made [his mum]$_{iobj}$ [a cake]$_{dobj}$.
3. Andrew made [his mum]$_{dobj}$ happy.

For English, we use one of the largest corpora available, the PukWAC (over 2 billion words, 30GB) (Baroni et al., 2009), which has been parsed by MaltParser (Nivre and Scholz, 2004). We extract all sentences with *to do* or *to make* (based on lemmata). The verb *to make* occurs in 2,13% of sentences, and the verb *to do* in 3,27% of sentences in the PukWAC corpus. Next, we exclude from consideration phrasal mono-transitives and select sentences where verb complements are nouns (Table 1).

For experiments in Dutch, we use the "Wikipedia Dump Of 2010" corpus, which is a part of Lassy Large corpus (159 million tokens), and is parsed by

LANG	# sent	# dobj (*to make*)	# dobj (*to do*)
EN	181,813,571	1,897,747	881,314
NL	8,639,837	15,510	6,197

Table 1: The number of sentences in English (EN) and Dutch (NL) corpora (the last two columns correspond to the number of sentences where direct objects are nouns).

the Alpino parser (Bouma et al., 2001). Unlike in English data, *to make* occurs here more often than *to do* (3,3% vs. 1%). This difference can be explained by the fact that *to do* is also an auxiliary verb in English which leads to more occurrences in total. Similarly to the English data set, phrasal monotransitives are filtered out. Finally, the sentences that contain either *to make* or *to do* from wiki01 up to wiki07 (19,847 sentences in total) have been selected for training and wiki08 (1,769 sentences in total) for testing. To be able to compare our results against the performance on English data, we sample a subset from PukWAC which is of the same size as Dutch data set and is referred to as *EN (sm)*.

To measure distributional similarity for the nearest neighbour method, we use first-order and second-order similarity based on Lin's information theoretic measure (Lin, 1998). For both languages, similarity scores have been derived given a subset of Wikipedia (276 million tokens for English and 114 million tokens for Dutch) using the DISCO API (Kolb, 2009).

2.3 Results

Table 2 and Table 3 summarize our results. When referring to similarity-based methods, the symbols **(f)** and **(s)** indicate first-order and second-order similarity. For the BoW models, ± 2 and ± 3 corresponds to the context length. The performance is measured by true positive rate (*TP*) per class, overall accuracy (*Acc*) and coverage (*Cov*). The former indicates in how many cases the correct class label (*make* or *do*) has been predicted, while the latter shows how many examples a system was able to classify. Coverage is especially indicative for LCS and semantic similarity approaches because they may fail to yield predictions. For these methods, we provide two evaluations. First, in order to be able to compare results against the BoW approach, we measure accuracy and coverage on all test examples. In such a case, if some direct objects occur very often in the test set and are classified correctly, accuracy scores will be boosted. Therefore, we also provide the second evaluation where we measure accuracy and coverage on (unique) test examples regardless of how frequent they are. This evaluation will give us a better insight into how well LCS and similarity-based methods work. Finally, we tested several settings for the LSC method and the results presented here are obtained for 20 clusters and 50 iterations. We remove stop words [2] but do not take any other preprocessing steps.

For both languages, it is more difficult to predict *to do* than *to make*, although the differences in performance on Dutch data (NL) are much smaller than on English data (EN (sm)). An interesting observation is that using second-order similarity slightly boosts performance for *to make* but is highly undesirable for predicting *to do* (decrease in accuracy for around 15%) in Dutch. This may be explained by the fact that the objects of *to do* are already very generic. Our findings on English data are that the similarity-based approach is more sensitive to the choice of aggregating over all words in the training set or selecting the nearest neighbour. In particular, we obtained better performance when choosing the nearest neighbour for *to do* but aggregating over all scores for *to make*. The results on Dutch and English data are in general not always comparable. In addition to the differences in performance of similarity-based methods, the BoW models work better for predicting *to do* in English but *to make* in Dutch.

As expected, similarity-based approaches yield higher coverage than LSC, although the latter is superior in terms of accuracy (in all cases but *to do* in English). Since LSC turned out to be the most computationally efficient method, we have also run it on larger subsets of the PukWAC data set, up to the entire corpus. We have not noticed any signifi-

[2] We use stop word lists for English and Dutch from `http://snowball.tartarus.org/algorithms/`.

LANG	Method	TP (*to make*)	Cov (*to make*)	TP (*to do*)	Cov (*to do*)	Acc (all)	Cov (all)
EN (all)	LSC	91.70	98.75	73.40	97.16	85.90	98.24
EN (sm)	LSC	89.81	90.00	75.81	86.70	86.91	89.30
	SMP (f)	84.89	98.82	69.89	95.14	81.78	98.03
	SMP (s)	82.92	98.82	55.65	95.14	77.27	98.03
	SMknn (f)	62.61	98.82	91.13	95.14	68.52	98.03
	SMknn (s)	4.36	98.82	99.46	95.14	24.07	98.03
	BoW ± 2	36.41	100	82.21	100	46.01	100
	BoW ± 3	32.26	100	84.10	100	43.13	100
NL	LSC	98.75	91.79	95.74	93.37	98.09	92.13
	SMP (f)	95.64	95.82	92.97	98.14	95.06	96.32
	SMP (s)	97.52	95.82	76.75	98.14	93.00	96.32
	SMknn (f)	94.14	95.82	92.97	98.14	93.89	96.32
	SMknn (s)	96.09	95.82	78.64	98.14	92.30	96.32
	BoW ± 2	89.34	100	61.19	100	83.44	100
	BoW ± 3	91.06	100	54.18	100	83.32	100

Table 2: True positive rate (*TP*, %), accuracy (*Acc*, %) and coverage (*Cov*, %) for the experiments on English (*EN*) and Dutch (*NL*) data.

LANG	Method	TP (*to make*)	Cov (*to make*)	TP (*to do*)	Cov (*to do*)	Acc (all)	Cov (all)
EN (sm)	LSC	80.88	77.12	52.60	74.76	73.73	76.51
	SMP (f)	73.17	97.29	45.99	90.78	66.49	95.60
	SMP (s)	77.00	97.29	33.69	90.78	66.36	95.60
	SMknn (f)	31.18	97.29	82.35	90.78	43.76	95.60
	SMknn (s)	4.36	98.82	98.93	90.78	25.76	95.60
NL	LSC	94.85	63.40	86.59	76.64	92.39	66.83
	SMP (f)	87.55	81.37	77.00	93.45	84.24	84.50
	SMP (s)	91.16	81.37	54.00	93.45	80.52	84.50
	SMknn (f)	80.72	81.37	76.00	93.45	79.66	84.50
	SMknn (s)	85.54	81.37	55.00	93.45	76.79	84.50

Table 3: True positive rate (*TP*, %), accuracy (*Acc*, %) and coverage (*Cov*, %) for the experiments on English (*EN*) and Dutch (*NL*) unique direct objects.

cant changes in performance; the results for the entire data set, EN (all), are given in the first row of Table 2. Table 3 shows the results for the methods using direct object information on unique objects, which gives a more realistic assessment of their performance. At closer inspection, we noticed that many non-classified cases in Dutch refer to compounds. For instance, *bluegrassmuziek (bluegrass music)* cannot be compared against known words in the training set. In order to cover such cases, existing methods may benefit from morphological analysis.

3 Conclusions

In order to predict the use of two often confused verbs, *to make* and *to do*, we have compared two methods to modeling selectional preferences against the bag-of-words approach. The BoW method is always outperformed by LCS and similarity-based approaches, although the differences in performance are much larger for *to do* in Dutch and for *to make* in English. In this study, we do not use any corpus of non-native speakers' errors and explore how well it is possible to predict one of two verbs provided that the context words have been chosen correctly. In the future work, we plan to label all incorrect uses of *to make* and *to do* and to correct them.

Acknowledgments

The author thanks anonymous reviewers for their valuable comments. This work is supported by a VICI grant number 277-80-002 by the Netherlands Organisation for Scientific Research (NWO).

References

Marco Baroni and Silvia Bernardini and Adriano Ferraresi and Eros Zanchetta. 2009. *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora.* Language Resources and Evaluation 43(3), pp. 209-226.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. *Alpino: Wide-coverage Computational Analysis of Dutch.* In Computational Linguistics in the Netherlands 2000. Enschede.

Josée A. Coenen, W. van Wiggen, and R. Bok-Bennema. 1979. *Leren van fouten: een analyse van de meest voorkomende Nederlandse taalfouten, die gemaakt worden door Marokkaanse, Turkse, Spaanse en Portugese kinderen.* Amsterdam: Stichting ABC, Contactorgaan voor de Innovatie van het Onderwijs.

Katrin Erk. 2007. *A simple, similarity-based model for selectional preferences.* In Proceedings of ACL 2007. Prague, Czech Republic, 2007.

Andrew R. Golding and Dan Roth. 1999. *A Winnow-Based Approach to Context-Sensitive Spelling Correction.* Machine Learning 34(1-3), pp. 107-130.

Peter Kolb. 2009. *Experiments on the difference between semantic similarity and relatedness.* In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09, Odense, Denmark, May 2009.

Dekang Lin. 1998. *Automatic Retrieval and Clustering of Similar Words.* In Proceedings of COLING-ACL 1998, Montreal.

Diana McCarthy and John Carroll. 2003. *Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences.* Computational Linguistics, 29(4), pp. 639-654.

Joakim Nivre and Mario Scholz. 2004. *Deterministic dependency parsing of English text.* In Proceedings of COLING 04.

Sebastian Padó, Ulrike Padó and Katrin Erk. 2007. *Flexible, Corpus-Based Modelling of Human Plausibility Judgements.* In Proceedings of EMNLP/CoNLL 2007. Prague, Czech Republic, pp. 400-409.

Mats Rooth, Stefan Riezler and Detlef Prescher. 1999. *Inducing a Semantically Annotated Lexicon via EM-Based Clustering.* In Proceedings of ACL 99.

Anna Rozovskaya and Dan Roth. 2010. *Generating Confusion Sets for Context-Sensitive Error Correction.* In Proceedings of EMNLP, pp. 961-970.

Sabine Schulte im Walde. 2010. *Comparing Computational Approaches to Selectional Preferences – Second-Order Co-Occurrence vs. Latent Semantic Clusters*. In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 1381–1388.

Beñat Zapirain, Eneko Agirre and Lluís Màrquez. 2009. *Generalizing over Lexical Features: Selectional Preferences for Semantic Role Classification.* In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore, pp. 73-76.

Harris Zellig. 1954. *Distributional Structure.* Word 10 (2/3), p. 146-62.