Using clustering to improve adjective selection in English adjective noun pairs – Rachele De Felice, Oxford University

This paper addresses the natural language generation problem of automatically selecting the correct collocation for adjective-noun pairs in English when two or more nearly synonymous alternatives are available but only one is appropriate, e.g. *tall(*high) man* vs. *high(*tall) status*. The problem is particularly evident in machine translation between languages which use a different number of adjectives to express the same property, e.g. Italian and English. For example, the property of height is expressed by *alto* in Italian, but by both *tall* and *high* in English: the translation system must choose between the two adjectives, selecting the collocationally appropriate one for the given noun. There is also a more complex instance of this problem, conceptual translation ambiguity: a noun may collocate with different adjectives in the two languages, i.e. adjectives which are not translations of each other. This makes literal translations nonsensical or difficult to understand, e.g. *nightmare* which is *brutto sogno* ('ugly dream') in Italian, but *bad dream* (It. 'cattivo sogno') in English.

We propose that Adj+N pairings can be predicted using cluster membership properties, based on the notion that adjective-noun collocations are not random, but follow certain patterns (e.g. *tall* and *short* with *man, woman, girl*; *fake* and *real* with *diamond, leather, gold*). We cluster nouns and assign adjective preferences to clusters rather than to individual nouns; in preliminary tests, our method achieves up to 78% accuracy. Furthermore, in solving the empirical problem we are presented with useful insights into the interaction of various parameters in clustering.

The approach so far has been tested on a set of 28 common English adjectives often found in complementary distribution (e.g. *strong tea* vs. **powerful tea*) and therefore susceptible to the translation ambiguities mentioned, such as *strong – powerful, fake – false, big – large – great.* To develop and test the system, a small corpus of British English was constructed from the British National Corpus. A set of nearly 200, 000 Adj+N pairs was extracted, comprising 2379 distinct nouns and their occurrences with one or more of the adjectives considered.

As this method rests on the notion that patterns of adjective collocation exist and are predictable for a given set of nouns, a way of grouping the nouns was needed which would yield sets exhibiting homogeneous behaviour with respect to adjective choice. Automated clustering was chosen as it avoids human bias while allowing underlying semantic properties to be captured, and enables the analysis of more data than is feasible manually. The intuition underlying the use of clustering is that if a set of nouns is observed to behave similarly in two contexts, in our case co-occurrence with a set of 250 nouns and 250 verbs, we can predict that they will behave similarly in a third context, too, namely co-occurrence with a given set of adjectives.

We assigned the nouns to one of 100 clusters and through pointwise mutual information and chi-square tests established that there is a correlation between cluster membership and adjective choice: significant scores for a cluster identify instances where an adjective occurs significantly more or less often than expected with nouns of that cluster, pointing to the nouns' strong preference or dispreference for that adjective. Cluster membership is shown to be a good indicator of adjective choice and on this premise we developed a simple model for the translation of Italian Adj+N pairs into English.

We assessed the model's performance against a baseline which used unigram probabilities only, i.e. the most frequent translation of the Italian adjective regardless of the co-occurring noun. The assumption is that with no other information, a translation system would use that value as a guide to lexical choice rather than choose randomly. 300 Italian Adj+N pairs were submitted for translation; we distinguished between Adj+N combinations not seen in our corpus, and those which were.

We find that the clustering model significantly outperforms the baseline in two tasks (percentage of correct results):

ClusteringModel – 75%
Baseline – 63%

Seen_data_only:
ClusteringModel – 78%
Baseline – 63%

Unseen_data_only:
ClusteringModel – 66.6%
Baseline – 64%

Despite the non-significant difference in the unseen data task, the results are positive as they show that the model overall is solid, and its two components can be used together without impairing its performance. Our model, unlike the baseline, successfully generates pairs like *heavy (*strong) infection*, *tall (*high) man*, *big (*great) breath*. We discuss the advantages the model has over the baseline in dealing with infrequent collocations, as well as possible causes for the drop in performance on the unseen data task.

Our results show that the clustering approach is viable and successful. The 75% success rate confirms the notion that clustering is a useful procedure for NLP applications, even when performed with basic constraints, as done here. These results were obtained using a small cluster set, with some large clusters susceptible to noise. It is hypothesised that more refined clustering – varying parameters such as number of clusters, feature window size, feature number/type, consideration of syntactic dependencies – could lead to a higher success rate. The usefulness of clustering for target-word selection in MT has been previously suggested (e.g. Dagan and Itai, 1994; Kikui, 1999). Our results reiterate these findings, and show their validity specifically for the domain of the translation of Adj+N pairs, which has so far received little attention. The clustering model is a helpful addition to NLG, and offers an improvement that is not only quantitative – as expressed by the significant difference in accuracy – but also qualitative, since, compared to the baseline, it extends the range of Adj+N pairs that are correctly translated. Lapata et al. (1999) suggested that it would prove difficult to generalise adjective preferences to unseen data, but we show that this is not the case, achieving an encouraging success rate of 66.6% in preliminary tests.

Finally, the paper also briefly surveys some recent work in the related field of 'generation-heavy' MT (Habash and Dorr, 2002), where most of the work is done at target language level – an approach found especially appropriate for language pairs with few parallel corpora and other resources used in 'traditional' stochastic MT. Our research supports this claim, showing that a successful translation module can be implemented in the absence of bilingual corpora, given sufficient target language information.