INTERVENTO

RACHELE DE FELICE – STEPHEN G. PULMAN
(Oxford University)

Using Clustering to Improve Adjective Selection in English Adjective-noun Pairs

0. PREAMBLE

One of the problems for natural language generation (*NLG*) regards the choice of the most appropriate collocate for a given word when two or more nearly synonymous alternatives are available (*e.g.* in English *tall* vs. *high*, *big* vs. *large*). These collocations often appear arbitrary and unpredictable. Data sparseness means that simple methods for predicting collocates, such as merely listing them, will have incomplete coverage. This paper presents a method for predicting the correct adjective based on cluster membership of the noun, where noun clusters are established on the basis of co-occurrence with high frequency verbs and nouns. The method achieves up to 78.6% accuracy, a substantial improvement over a simple frequency baseline, and is the basis for some considerations on the usefulness of clustering in Natural Language Processing (*NLP*) applications.

1. INTRODUCTION

Natural language generation of English often runs into problems when the language offers two or more nearly synonymous alternatives but only one is appropriate in the given context – cf. for example *tall(*high) man* vs. *high(*tall) status*, or *bake a cake* vs. *roast a joint*. This issue is exacerbated in machine translation (*MT*) into English from languages, such as Italian, whose lexicon does not map directly onto the English one. One aspect of this issue is the problem addressed in this paper: a 'mis-match' between adjective-noun collocations in the two languages, which manifests itself in two ways. In the simpler case, the two languages express the same property through a different number of adjectives, not all of which usually collocate with a given noun: *e.g.* the quality of height is expressed by *alto* in Italian, but by both *tall* and *high* in English. This requires the translation system to operate a choice between the two adjectives, ideally selecting the collocationally more appropriate one. The more complex case involves what is known as conceptual translation ambiguity: a given noun collocates with

different adjectives in the two languages, that is, adjectives which are not translations of each other. This makes literal translations nonsensical or difficult to understand: *e.g.* a *nightmare* is *brutto sogno* 'ugly dream' in Italian, but *bad dream* (It. *cattivo sogno*) in English. While it might be feasible to just list the correct adjective for frequently occurring combinations, complete coverage would never be possible because of data sparseness. For example, of the English *Adj+N* combinations described below, only 73% occur in the British National Corpus (*BNC*, cf. Burnard, 1995), and each of the adjectives and nouns is individually of high frequency. Estimates for unseen data would have to be recreated using smoothing or similar methods (*e.g.* using unigram estimates). Dagan *et al.* (1997) suggest that similarity based language models are preferable to smoothing and other 'classical' back-off methods in dealing with data sparseness[1], and this is the approach we have taken here.

This paper proposes that it is possible to predict *Adj+N* pairings, based on cluster membership, thereby eliminating the need for all of them to be stored by a system. On the basis of the notion that 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 present a method whereby nouns are clustered and adjective preferences are assigned to clusters rather than to individual nouns. In the testing we have carried out so far, our method achieves up to 78.6% accuracy.

The research we present here has the immediate, technical, aim of improving the accuracy of the generation of translated *Adj+N* pairs, but it also has the broader aim of investigating the contribution of clustering-based methods to the field. Clustering, it seems to us, is a powerful *NLP* technique as it is a way to efficiently harness intuitions we have about language, supporting them through the analysis of millions of words, which can only be done computationally. The translation task described, then, is a convenient, small-scale exercise through which we can assess the value of clustering.

2. Methodology

In keeping with our desire to start on a small scale, we selected only 28 English adjectives for testing. These are presented in Table 1 together with their Italian counterparts.

[1] "Similarity-based models assume that if word w'_1 is 'similar' to word w_1, then w'_1 can yield information about the probability of unseen word pairs involving w_1" (Dagan *et al.*, 1997). The experiments described in that paper do not, however, involve clustering techniques.

Table 1. *Adjectives considered in this study*

Bad	Cattivo
Beautiful	Bello
Big	Grande, grosso
Fake	Falso
False	Falso
Fast	Veloce
Fat	Grasso
Good	Buono
Great	Grande
Handsome	Bello
Heavy	Pesante
High	Alto
Large	Grande
Light	Leggero
Little	Piccolo
Low	Basso
Nice	Bello
Powerful	Potente
Quick	Veloce
Real	Vero
Short	Corto, basso
Small	Piccolo
Strong	Forte
Tall	Alto
Thick	Spesso
Thin	Sottile, magro
True	Vero
Ugly	Brutto

They were chosen according to the following criteria: being often found in complementary distribution (*e.g. strong tea* vs. **powerful tea*), they are potentially very susceptible to the translation ambiguities described above; they are basic, frequent adjectives of the language[2]; and they are the most common, neutral level of the quality they refer to, so for example we have *fat* but not *obese, beautiful* but not *gorgeous*, and so on. It was felt it was best to begin with a small, simple set

[2] For instance, many of them belong to a set of 'core' adjectives present in almost all languages of the world, referring to certain general properties such as dimension or value (see *e.g.* Dixon, 1982).

of words, and then experiment with other types of adjectives once this method was found to be successful. It can be noted from the table that not all the adjectives appear involved in ambiguities: for example, while *grande* maps to *big, large,* and *great*, *forte* only maps to *strong*. However, our task takes account of both kinds of ambiguity described above, and therefore needs to also cover cases such as the collocational distinctions between, say, *strong* and *powerful*, hence the inclusion of these adjectives in the table.

To develop and test the system, a small corpus of British English was constructed from the *BNC*. 2379 distinct nouns were selected, chosen because they appeared in the *BNC* to the immediate right of one of the adjectives under examination more than 5 times (to minimize noise caused by *ad-hoc* usage). Manual post-editing was also carried out to minimize the risk of false positives, i.e. those instances where the adjective did not in fact modify the noun immediately following it. This approach, which may seem rather restrictive and may potentially have led to the loss of some useful data, was deemed the most appropriate for a small-scale study such as this, as it eliminated the need for syntactic or phrasal analysis.

This procedure yielded a subcorpus consisting of nearly 400,000 words, including nouns from a diverse range of domains: family terms (*mother, father, brother*); plants and animals (*mammal, flower, dog, mouse*); concrete items (*plate, mug, table, cupboard*); abstract qualities (*imagination, courage, ability*); scientific and mathematical properties (*specificity, average, weight*); business terms (*dividends, investor, profit*); and many more.

Central to our method is the notion that patterns exist in language and can be predicted. In this particular case, we are concerned with patterns of adjective collocation, and how they can be predicted for a given set of nouns. Therefore, the nouns selected had to be grouped in some way, ideally yielding sets which exhibit homogeneous behaviour with respect to adjective choice. The fact that group membership can potentially predict adjective selection has the additional benefit of allowing the predictions to be extended to novel words: their adjective preferences can be determined on the basis of those of the group they are most similar to. The first approach to this problem of group creation followed human intuition, and assumed that clusters could be easily constructed on the basis of surface semantic properties, *e.g.* 'family relations' comprising *father, mother, son, daughter*. This was ruled out when it was noticed that even nouns which are apparently closely related do not in fact share adjective preferences: cf. *real mother*, *real father*, vs. *true son*, *true daughter*. *WordNet* synsets were also considered as a source of semantically related sets of words, but a quick overview of the data revealed that they, too, were not appropriate for this task. Not all members of a synset necessarily select the same adjective given a choice of two or more, while there are other words that do choose the same adjective and are arguably related to each other, but not in a way captured by *WordNet* (*e.g.* in it *gambler* is not related to *smoker* and *drinker*, yet they all collocate with adjectives such as *heavy*;

leather is not related to *pearls* and *diamonds*, but all three prefer *fake* over *false* and *real* over *true*). In future trials, we plan to investigate and compare performance using clusters constructed on the basis of these initially discarded approaches.

A good alternative to manually-created ontologies is offered by vectors and clustering. Vectors are based on the principle that words with similar meanings are often used in similar contexts and have similar distributions (Miller and Charles, 1991). Of course, this idea is not an innovation of computational linguistics: for decades, budding linguists have been told that "You shall know the meaning of a word by the company it keeps" (Firth, 1957: 179). More recently, Cruse (1986) discussed the "contextual approach" to meaning, which aims to derive the semantic properties of a lexical item from the relations it contracts with actual and potential linguistic contexts. We can also test this informally. Given the following sentences with the non-existent word *birk*:

(1) I ate a plateful of birk.
(2) Birk is best cooked in copper pans.
(3) She had never tasted birk before.

one will probably infer on the basis of the things one can and cannot do with it – the words making up its context – that this mysterious *birk* is a foodstuff. *Birk* is not synonymous with *apple, pear*, or *chicken*, but it is interchangeable with these words, suggesting they share an underlying semantic property.

Vectors formalise this intuition about language by defining the words analysed (target words) in terms of the words making up their context (features, usually nouns and verbs). An algorithm keeps track of which features occur around the target word, and how often, so that each target word is identified by a vector. It follows that two words that appear in similar contexts will have vectors that are more similar to each other than two words that do not. The resulting vectors are then usually clustered according to the similarity between vectors. Vectors, then, act as representations of word senses. In fact, it can be argued that this means of representing word senses is more useful than a dictionary or encyclopaedia entry, because semantic distance or relatedness between words can be easily and quickly calculated, as well as visualized graphically. This property makes the vector space model a commonly used tool in tasks such as word sense disambiguation and sense discrimination. It is an empirical approach which fits well with the results- rather than theory-oriented perspective of computational linguistics. The use of vectors as word senses has been proposed by, among others, Schütze (1998), who states that "a sense is a group of contextually similar occurrences of a word" (Schütze, 1998: 99). It is the syntactic environment of the target words, then, rather than obvious semantic properties, that becomes the basis of clustering: it is an indirect manifestation of underlying semantic characteristics. Intuitively, too, this makes sense: if clusters group words which are to some degree mutually interchangeable, that is, they can perform the same task in the sentence, this interchangeability can only derive from the words sharing some property.

Automatically induced vectors and clusters, then, are an ideal way of capturing this underlying shared property. An important advantage unsupervised clustering methods have over manual ones is that they are not susceptible to human biases and preconceptions about how data 'should behave': they only take into account what the data actually does. Furthermore, clustering through machine learning allows large amounts of data to be analysed and clustered, far more than could be ever examined by a human, allowing for more comprehensive and potentially more insightful results. In light of all this, it was decided to arrange the nouns under study using automatically induced clustering. The ultimate result will not be too different from *WordNet*, in that we will have groups of mutually interchangeable words that are somehow related to each other, but these will have been determined using different criteria and methods.

The clustering was performed by a k-means algorithm[3], using as features the 250 most frequent verbs and 250 most frequent nouns in the *BNC*, including any target words that might occur in the latter list, but excluding a set of stop words. The vectors were constructed by considering co-occurrences of the features within a window of 10 words either side of the target word[4]. The clustering results were then evaluated to determine whether there is a significant correlation between cluster membership and adjective choice: if not, there is clearly little likelihood of cluster membership being a good predictor of adjective choice.

3. Cluster-adjective correlation

Various values of K, from 50 to 500, were used in performing the clustering. The set comprising 100 clusters was chosen to develop the system, as it offered the best trade-off between internal coherence, size, and informativeness. Having a small number of clusters means that each cluster is very large and rather noisy. Conversely, while sets with many clusters have refined, tightly coherent clusters, these will be very sparsely populated, often having no more than two or three members. The set with 100 clusters proved to be the least susceptible to these problems, displaying clusters with little noise and a good number of members – only one has two members only.

[3] Once it has been decided how many clusters are needed (k), the algorithm works by selecting k data points for k clusters as centers of the clusters. The remaining data points are assigned to the cluster they are closest to. It then calculates the mean for each cluster, and data points can change cluster membership if there is another cluster they are closer to after this calculation. The process is iterated until there are no more variations in cluster membership.

[4] There are a variety of opinions in the literature as to what the optimum window size is, with numbers ranging from 5 to 30; both small and big sizes have their advantages, offering either words which are more closely related to the target word (small) or bringing in more context and therefore more content (big); see *e.g.* Aston and Burnard (1998), Marx and Dagan (2002), Pedersen *et al.* (2005). Our choice here is to find a balance between the various factors.

Correlation between clusters and adjectives was established using both pointwise mutual information (*MI*) and χ^2 tests. These measures were chosen as they point out instances of the adjective occurring at a significantly higher or lower frequency than expected. A significant score indicates a significant relationship, which can be one of either strong preference – if the adjective occurs significantly more frequently than expected – or strong dislike – if the actual frequency is significantly lower than expected. χ^2 scores were deemed to be significant for $p \leq 0.001$. Each cluster was assigned a score for each adjective, giving 28 scores for each cluster. These were calculated by comparing the expected frequencies of the given adjective for that cluster, derived from the overall frequency of the adjective and the size of the cluster, to the observed frequency of the adjective for that cluster. *MI* was used as well, to complement and reinforce the information offered by the χ^2 test, *e.g.* when the latter's scores were not very significant. *MI* is an appropriate choice as it answers a similar question to that posed by the χ^2 test: is there a significant relation between a given word and its immediate lexical surroundings? This is measured by comparing the probability of observing two given words together (joint probability) with the probabilities of observing each of the words independently (Church and Hanks, 1989). If one of the words 'forces' certain terms rather than others to appear near it, the joint probability will be greater than the individual, or chance, probabilities of those two words, and *MI* will be > 0. If there is no special relation between the words, joint probability and chance probability will be about the same; *MI* will be ≈ 0. If the two words are in complementary distribution, i.e. they show a strong aversion to occurring together, the joint probability will be less than the chance probability, and *MI* will be < 0. It will be immediately noted that, unlike the χ^2 test, for *MI* the larger the score, the more significant the correlation between two words is. Church and Hanks (1989) suggest that a score of 3 or above signals a significant correlation between words, and we adopt this guideline here. The *MI* score for each cluster derives from the sum of all the pointwise *MI* scores for the nouns of that cluster, so for significant relationships, those where most words in the cluster have a significant correlation with a given adjective, it is likely to be much higher than 3.

The validity of these scores as predictors for adjective choice, and their interpretation, will be illustrated with two examples, with reference to only two clusters for reasons of space and clarity. Table 2 shows some data for cluster 17, which includes 4426 *Adj+N* occurrences.

Table 2. *Sample of the scores for cluster 17*

	Obs.	**MI**	**P**
TALL	0	--	0.000263
HIGH	1224	70.82	*0*
BIG	12	2.09	1.22^{-53}
LARGE	*1341*	*48.4*	*0*

The first column shows the observed frequencies of the four adjectives listed for that cluster, the second column the *MI* scores, and the third the results of the χ^2 test. *Tall* does not receive an *MI* score due to there being 0 occurrences of the adjective for that cluster. The p-scores tell us that all four adjectives have some significant relation with the nouns of the cluster, and the observed frequencies, together with the *MI* scores, suggest the direction of these relations. It can be inferred that the nouns of cluster 17 – shown below – have a strong preference for *high* and *large*, and are very unlikely to co-occur with *tall* and *big*. Looking at the actual composition of the cluster shows that these predictions are confirmed: one says, for example, *high average* and *high margins* rather than *tall average* and *tall margins*, and *large quantity* and *large fraction* rather than *big quantity* and *big fraction*.

```
cluster 17.100
adjustments availability average calculation
circulation comparison component concentration
consumption conversion coverage duration
efficiency estimate excess exposure extract
fraction grade growth increments index
indication indicator margin medium output
percentage population proportion quantity
rating ratio reduction sample threshold
timing volume yield
```

Table 3. offers an example with a different pair of adjectives, *small – little*, and another cluster, cluster 53, consisting of 766 occurrences.

Table 3. *Sample of the scores for cluster 53*

	Obs.	**MI**	**P**
SMALL	66	15.68	0.042534
LITTLE	*373*	*99.2*	*0*

We can see that the scores of the significance tests point overwhelmingly in the direction of a preference for *little* over *small*. Indeed, the prediction is confirmed when we consider the cluster's elements. One usually says *little darling, little nod*, and so on, rather than *small darling* or *small nod*.

```
cluster 53.100
astonishment bastard bugger cheek chuckle
darling eyebrows frown gasp grimace
grin groan growl grunt gulp hug hurry
kiss moan murmur nod scream shit
shiver shout shrug shudder sigh sniff snort tears
```

These brief examples show that cluster membership correlates well with adjective preferences, and suggest that it can be used to predict adjective choice when two or more alternatives are available. They also give an idea of the internal homogeneity of the clusters: it can be seen that their components, though not synonyms of each other, do seem to belong to broadly the same domain of discourse, and occupy similar positions within a sentence. This suggests that the parameters we have chosen to perform our clustering are solid ones, though we plan to experiment with variations on one or more of the parameters, such as number of clusters, window size, or co-occurrence with other parts of speech.

The information discussed above was used to develop a simple model for the translation of *Adj+N* pairs, with cluster membership as the guiding criterion for selecting English translations of Italian adjectives. The model first establishes what cluster the noun under examination belongs to. It then looks at the scores for the relevant cluster for each of the possible translations of the adjective (*e.g. big, large*, and *great* for It. *grande*), choosing the one which presents a significant p-score together with a high *MI* score. Use of the *MI* score as well is essential as the model does not have access to the actual frequency counts for each adjective, and a significant p-score by itself is not enough to decide whether the adjective is strongly preferred or dispreferred, since it can mean either a significantly higher- or lower- than-expected frequency. A variant of the model is planned, where the frequency information is available to the model, and *MI* scores are not considered.

4. Testing the model

The performance of the model was assessed against a baseline which used unigram probabilities only, that is, the most frequently occurring translation of the Italian adjective, regardless of the noun co-occurring with it. The assumption is that in the absence of any other information, a translation system would use that value as a guide to lexical choice rather than just choosing randomly. 450 Italian *Adj+N* pairs were submitted for translation; to avoid introducing extraneous complications, it was assumed the nouns were unambiguously and correctly translated, leaving adjective choice as the only task to focus on. The pairs were selected so as to ensure a wide number of clusters and adjectives were represented, and a sizeable amount of 'unseen' data was included. Because of the way we had formed the clusters, the notion of 'unseen' data is not straightforward here: since all the noun clusters were formed without reference to adjective co-occurrence, from one point of view all of our *Adj+N* test examples were unseen. However, some of the noun occurrences in our corpus were in an adjectival modification context, and so strictly speaking these were not literally unseen, and some may even have been seen in the context of the adjectives we were interested in. It is difficult to see how this could have affected our results, however, since the likelihood of a verb+noun combination is unlikely to be affected by any adjectives

present: selectional restrictions, for example, hold between a verb and the head noun of any complement NPs. Nevertheless, we separated out the two classes of noun: those not seen in any *Adj* combination, and those that had, and to our surprise, there is a difference in accuracy, as discussed below.

Evaluating the accuracy of a translation is generally agreed to be a demanding task (see *e.g*. Knight and Marcu, 2005; Habash *et al*., 2003), and in dealing with adjectives, it is especially difficult to decide between 'right' and 'wrong' answers. Adjective collocation, unlike verb-argument relations, is an area where traditional grammars and strong intuitions seldom offer clear guidelines as to what is correct and what is not. It is hard to explain why one says that *fake promise* and *big amount* are 'wrong' and *false promise* and *large amount* are 'right', when all are perfectly intelligible. It is likely that judgements about the appropriateness of adjective selection come mostly through frequency of use, not from any grammar book rules. We tend to say that *false promise* is better than *fake promise*, and sounds more natural, because that is the most common version of the phrase: patterns and frequency rather than strict grammaticality considerations determine acceptability (see *e.g*. Biber *et al*., 1998; Jurafsky and Martin, 2000: 598-9). Accordingly, frequency of use is the principle we followed in evaluating 'right' and 'wrong' answers in the test results, and the principle is followed very strictly. This means that even an output that is considered acceptable, and is perfectly intelligible, such as *little cave*, is counted as wrong since frequency counts seem to prefer *small cave*. This rigid approach was adopted to allow for uniform treatment of all cases and to avoid having to distinguish between various degrees of 'wrongness' (*e.g*. *ugly dream* is clearly 'more wrong' than *fake promise*, but the model does not make this distinction and simply marks both of them as incorrect). For the purposes of this experiment, the 'correct' English *Adj+N* pairing was deemed to be the one yielding more results in a web search, supported by findings in the *BNC*[5].

Table 4 summarises the results of the task: the clustering-based model, with 74% of correct *Adj+N* pairings, significantly outperforms the baseline (61%).

Table 4. *Translation task result*

	Baseline	**Clustering**
Correct	274 (61%)	333 (74%)
Incorrect	176 (39%)	117 (26%)

[5] The usefulness and reliability of web counts for NLP is often discussed (Keller and Lapata, 2003; Lapata and Keller, 2004); it was decided that for the purpose of obtaining information about relative frequency of data, the exact reliability of counts was not an issue, since all that was of interest was which of two or more alternatives is more frequent, not its exact frequency. We did of course try to minimise noise in these counts, and restricted our search to UK sites only.

We also compared performance when considering only those pairs which had occurred in the corpus (seen), and those which had not (unseen), to establish how far the usefulness of clustering stretches. A potential advantage of using clusters lies in the way they can be used for predictions, and for extending known patterns to novel data. By isolating the set of unseen data, we have a clear picture of how well the model performs when confronted with novel data, and how much, if any, of an improvement is introduced by clustering. Table 5 presents these results, which are rather encouraging.

Table 5. *Comparison of performance on seen vs. unseen data*

	Baseline	**Clustering**
Seen _correct	186 (62%)	236 (78.6%)
Seen _incorrect	114 (38%)	64 (21.4%)
Unseen _correct	88 (58.6%)	97 (65%)
Unseen _incorrect	62 (4.1.4%)	53 (35%)

In the case of seen data, the performance gap between the two models widens in favour of our model, which achieves 78.6% accuracy, a statistically significant improvement over the baseline's 62%. On the unseen data task, however, the two models, while still showing some degree of success, do not perform as well. Although the clustering model achieves an improvement of over 6% on the baseline, this difference is not statistically significant; we address possible causes for this below. The results are nevertheless very positive as they show that the clustering-based model as a whole is a solid one, and its two components – seen and unseen – can be used together without seriously impairing performance.

5. Discussion

Our results show that introducing a clustering-based component is a viable and successful approach. The 74% success rate is important because it confirms the notion that clustering is a useful procedure for *NLP* applications, even when performed with very basic constraints, as in this case. Given the success of this preliminary investigation, it is reasonable to assume that more refined clustering, whose output is potentially less susceptible to noise, could lead to a higher success rate. The usefulness of clustering for target-word selection in *MT* has been previously suggested (Dagan and Itai, 1994; Kikui, 1999; among others). Our

results reiterate these findings, as well as showing their validity specifically for the domain of the translation of *Adj+N* pairs, which has so far received little attention.

The main advantage this model presents over the baseline is that it can correctly identify those cases where the preferred translation of the adjective is not the most frequent one, due to the collocational restrictions of the noun. For example, *heavy infection* in Italian is *forte infezione*. The baseline here would default its adjective choice to *strong*, since that is the most frequent translation of the Italian adjective. The clustering model however knows that Italian *forte* can map to English *powerful* or *heavy* as well as *strong*. Therefore it compares the scores for all three possible translations for the cluster *infection* is in before making its selection, which is the correct one, *heavy*. Similarly for cases such as *tall man* (uomo alto) or *bad dream* (brutto sogno): the information the clustering model has about the closer relationship between these nouns and adjectives allows it to override the default translation of the adjectives as *high* and *ugly*, respectively.

As is clear from the test results, both models make mistakes. Where the clustering model makes mistakes, generally the baseline does, too. A closer analysis of the clustering model's errors reveals that they mostly involve cases where some elements of the cluster have a very strong correlation with the other, incorrect, adjective, and that strong preference is extended to the whole cluster. An example illustrative of this issue is *grande problema* (big problem) which the clustering model translates as *large problem*. The cluster *problem* is in is very large and includes several abstract nouns which do indeed prefer to collocate with *large* rather than *big*, leading to the significant score for this adjective for the cluster. Therefore despite the fact that *big problem* occurs in the corpus, its preference is overridden by the cluster's behaviour as a whole. This suggests that smaller, more refined clusters could have a positive effect on further improving the performance of the clustering model against the baseline, provided of course that the improvement in loss of noise is not counterbalanced by a restriction in cluster size that would make statistical analysis unreliable.

Conversely, there are several instances in which the baseline makes a mistake but the clustering model doesn't. These involve mostly less frequent adjectives, and point to the real advantage offered by the clustering model: the improvement introduced is not only quantitative, as expressed by the significant difference in the success rate, but also qualitative in that, compared to the baseline, the range of Adj+N pairs that are correctly translated is extended.

With regard to the unseen data task, we can see that the clustering model does not compare unfavourably to the baseline on that, either, although its accuracy was noticeably lower than that obtained in the seen data task. An analysis of the unseen data's error patterns revealed similar trends to those observed for the seen data, partly confirming our expectation that there should be little difference between the two groups of data. For example, as for the seen data, deciding between *big, large*, and *great* is often difficult for both models. Where the base-

line failed and the clustering model succeeded, instead, was, as before, in cases requiring the choice of a less frequent adjective. Examples of unseen pairs translated correctly by the clustering model but not the baseline are *grande respiro* (big breath; baseline selects *great*) and *corridore veloce* (fast runner; baseline selects *quick*). On the contrary, mistakes made by the clustering model but not the baseline – such as outputting *true consciousness* instead of *real consciousness* for the It. *vera consapevolezza* – are, like in the case of the seen data, due to the composition of the cluster the nouns belong to. *Consciousness*, for example, is in a cluster with *religion* and *believer*, which have a strong preference for *true*, and this preference is extended by the model to all the cluster's members. However, the unseen task presents an additional complication, which is in all likelihood the cause of the lower accuracy rate: some of the unseen data used in the task consists of a noun paired with an adjective for which there is no relevant information for the noun's cluster at all, in which case the χ^2 test score essentially mirrors the frequency of the adjective, offering no advantage over the baseline model. Indeed, if we consider only those instances of the unseen data for which there is some information for the adjective in the cluster under consideration, the performance gap between the clustering model and the baseline increases again, by almost 9%.

The above considerations suggest that, as anticipated, there are many benefits to be had from the use of clustering techniques in *NLP*, and more specifically in our case in *NLG* and *MT*. Our preliminary tests show that clusters are indeed a useful source of generalisations, allowing information known about some of its elements to be extended to other, previously unknown ones. This is invaluable in circumstances where the data one is working with is unseen, or very infrequent, so that little reliable information about it is available. Clustering offers a viable solution to overcoming these instances of data sparseness. Furthermore, in the context of *MT*, a model such as ours which shifts the bulk of the work on the target language rather than the source one could prove useful if the source language is one for which there are few resources available, *e.g.* no aligned *corpora* (cf. also the related work mentioned below). Of course, due to the small-scale nature of our investigations, and the many parameters which can be varied in the set up, further experimenting is needed to establish more firmly how to best make use of the contributions offered by clustering. We also plan to compare the model against other methods and baselines, computed in different ways.

It can be argued that clustering techniques can be viewed as a felicitous blend of linguistic intuitions and statistical methods. Improvement over the baseline did not require much effort in terms of understanding linguistic theories, or coming up with new ones, but it did require an efficient use of corpus resources. The improved performance of the clustering-based model is due to its exploitation of the facts of the language, that is, the way certain sets of words show a similar behaviour in various linguistic environments. This linguistic intuition, however, could not have been adequately tested by manual means alone: large amounts of data have to be analysed for reliable regularities to be discovered. Statistical tech-

niques such as clustering and χ^2 tests enable us to systematize the relations between the words of the language, and their behaviour, so as to enhance simpler 'dumb' models, without the need to teach them intricate linguistic theories. The small enhancements introduced by our model were sufficient to achieve a significant improvement over an 'ignorant' raw-frequency baseline. This is a crucial result: it goes to support the views of those, such as Dagan and Itai (1994), or Hatzivassiloglou (1996), to name just a few, who believe that successful computational linguistics cannot do without some form of linguistically-motivated input. We do need linguistics, even if we have statistics, because it allows us to get (for example) more accurate translations, more often. The use of statistics is of course central to the field, and to our model; but we believe we can derive the greatest advantages from these figures when they are used to represent a linguistic scenario accurately.

Bennett (2003: 145) suggests that applying methods of contrastive linguistics and second language acquisition to *MT* might prove successful, "after all, making generalizations that help a learner avoid mistakes in a second language is not all that different from writing rules that enable a *MT* system to produce adequate translations". The linguistic knowledge in our model was used in this spirit. Human learners improve their knowledge of the foreign language by learning when to generalise the notions they have acquired, and how to use analogy to their advantage. Using clustering as the foundation for adjective selection is a way of using analogy, especially for unseen data: if these two nouns behave in the same way in a first and second situation, it is legitimate to assume they will also do so in a third. This method, although not fail-safe, does lead to a fairly good performance and opens up many possibilities for further refinement.

6. Related work

There is not a great amount of work, in *MT* or in *NLG* as whole, concerned with *Adj+N* pairs (a notable exception is Lapata *et al*., 1999). There is also very little recent work on the language pair Italian-English. Furthermore, unlike *MT* models described in the literature, our model focuses on the treatment of *Adj+N* bigrams rather than entire sentences, which means it does not need to take into account information about syntax or word order. Therefore, our results cannot be directly compared to other, recent work in the field. Adding to this problem is the fact that, as discussed above, evaluating *MT* is notoriously difficult. So, while we can claim that the clustering model performs well in the tasks it was set, we cannot compare its 74% success rate directly against any other previous work.

Our results can however be viewed in the broader context of *NLG*, where since the presentation of Nitrogen, now known as HALogen (see *e.g.* Langkilde

and Knight, 1998; Langkilde-Geary, 2002), considerable attention has been paid to systems which shift the burden of the work to the generation component and target language knowledge. In *MT*, this means that the analysis and translation modules can output any number of potential translations of the source text, and the generation module, using statistical and symbolic information, will ultimately select the one to output. Habash *et al*. (2003) offer a good overview of the importance of these 'hybrid' *NLG* systems; the concept underlying our project, namely giving the generation end of *MT* more decisional power, is in the same spirit as these systems. Inkpen and Hirst (2003) present a "natural language generation system capable of distinguishing between near-synonyms". Their system, Xenon, uses a variety of sources – including χ^2 test scores and *MI* – to derive information about collocation patterns and stylistic nuances of English. It is tested on a translation task from French into English, and in generation of English from English (an English sentence is transformed into its interlingua representation and re-generated). For the latter task, the accuracy score is, unsurprisingly, very high (98 – 100%). For the translation task, accuracy is between 64% (baseline 35.7%) and 78% (baseline 76%). The translation scores are not directly comparable to ours because of some differences between the two projects: Xenon uses French, not Italian, it involves several POS, and is concerned with stylistic nuances as well as general collocation trends (for example, it contrasts *thin* with *lean, scrawny, slender, slim, willowy*). However, it will be immediately noted that despite the more complex set-up of Xenon, and the larger amount of sources of information about near-synonyms it uses, its performance is in the same league as that of our much simpler system, even allowing for the fact that Xenon had to translate entire sentences rather than bigrams and was working through an interlingua.

Habash *et al*. (2003) describe another hybrid *NLG* system, Lexogen, which is concerned with lexical choice from an argument- and thematic-role-selection perspective. Lexogen was tested on a Chinese-English *MT* task, evaluated on a scale from 1 (lowest) to 5 (highest): its average scores for both accuracy and fluency were slightly above 3. Allowing for some differences between this test and our research, it can be said that the paper shows that this hybrid approach is worth pursuing, and our work compares favourably with these more complex systems. The authors also call this approach 'generation-heavy' *MT* (Habash and Dorr, 2002), since a lot of the work is done at target language level, and suggest it is especially appropriate for those language pairs where there is a scarcity of parallel text and other similar resources typically used for 'traditional' stochastic *MT*. Our research supports this claim, showing that a successful translation module can indeed be implemented in the absence of aligned bilingual corpora, given sufficient target language information. This brief and by no means complete overview of the field suggests that our basic model is not underperforming compared to other, better-established models, and points to areas one can focus on to improve its performance.

7. Conclusion

We described a problem affecting *MT* from Italian into English, involving the selection of correct translations for adjectives. A solution to the problem was proposed, through which we also aimed to evaluate the usefulness of clustering in some domains of *NLP*. Our solution involved the clustering of English nouns to exploit the intuition that that nouns in the same cluster would co-occur with the same adjectives. Using pointwise mutual information and the χ^2 test, significant correlations were found between clusters and adjective preferences. This property was used to develop a simple *MT* model which was tested on a translation task involving both seen and unseen pairs of adjectives and nouns. This model achieved up to 78.6% accuracy compared to a baseline model using only frequency information (61% accuracy). Our results prove the viability of the clustering-based model, and the testing process offered useful methodological insights into the best approach to clustering and the determination of correlation between clusters and adjectives. Lapata *et al*. (1999) suggested that it would prove very difficult to generalise adjective preferences to unseen data, but this work shows that this is not in fact the case, achieving an encouraging success rate of 65% on a preliminary test on a small sample of unseen data.

REFERENCES

Aston G., Burnard L., 1998, *The BNC Handbook: Exploring the British National Corpus with SARA,* Edinburgh, Edinburgh University Press.

Bennett P., 2003, The Relevance of Linguistics for Machine Translation, in H. Somers (ed.) *Computers and Translation: a Translator's Guide*, Amsterdam, John Benjamins: 143-160.

Biber D., Conrad S., Reppen R., 1998, *Corpus Linguistics: Investigating Language Structure and use*, Cambridge, Cambridge University Press.

Burnard L., 1995, *The Users Reference Guide for the British National Corpus*, Oxford, British National Corpus Consortium, Oxford University Computing Service.

Church K., Hanks P., 1989, Word Association Norms, Mutual Information, and Lexicography, in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver: 76-83.

Cruse D. A., 1986, *Lexical Semantics*, Cambridge, Cambridge University Press.

Dagan I., Itai A., 1994, Word Sense Disambiguation Using a Second Language Monolingual Corpus, in "Computational Linguistics", 20.4: 563-596.

Dagan I., Lee L., Pereira F., 1997, Similarity-Based Methods for Word-Sense Disambiguation, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid: 56-63.

Dixon R., 1982, Where Have all the Adjectives Gone?, in *Where Have All the Adjectives Gone?: and other Essays in Semantics and Syntax*, Berlin, Mouton: 1-62.

Firth J. R., 1957, A Synopsis of Linguistic Theory, in F. R. Palmer (ed.), 1968, *Selected Papers of J.R. Firth 1952-9*, Harlow, Longmans: 168-205.

Habash N., Dorr B., 2002, Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation, in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, Tiburon, California: 84-93.

Habash N., Dorr B., Traum D., 2003, Hybrid Natural Language Generation from Lexical Conceptual Structures, in "Machine Translation", 18.2: 81-128.

Hatzivassiloglou V., 1996, Do We Need Linguistics when We Have Statistics? A Comparative Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System, in *Proceedings of the ACL – 94 Workshop on Combining Symbolic and Statistic Approaches to Language*, Cambridge, Mass.

Inkpen D., Hirst G., 2003, Near-synonym Choice in Natural Language Generation, in *Proceedings of the International Conference in Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria: 204-211.

Jurafsky D., Martin J., 2000, *Speech and Language Processing*, London, Prentice Hall.

Keller F., Lapata M., 2003, Using the Web to Obtain Frequencies for Unseen Bigrams, in "Computational Linguistics", 29.3: 459-484.

Kikui G., 1999, *Resolving Translation Ambiguity Using Non-Parallel Bilingual Corpora*, in *Proceedings of the ACL – 99 Workshop on Unsupervised Learning in NLP*, College Park, Maryland.

Knight K., Marcu D., 2005, Machine Translation in the Year 2004, in *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICAS-SP)*, Philadelphia, Pennsylvania.

Langkilde I., Knight K., 1998, Generation that Exploits Corpus-Based Statistical Knowledge, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal: 704-710.

Langkilde-Geary I., 2002, An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator, in *Proceedings of the International Language Generation Conference*, New York.

Lapata M., McDonald S., Keller F., 1999, Determinants of Adjective-Noun Plausibility, in *Proceedings of the 9th Conference of the European chapter of the Association for Computational Linguistics*, Bergen: 30-36.

Lapata M., Keller F., 2004, The Web as Baseline: Evaluating the Performance of Unsupervised Web-Based Models for a Range of NLP Tasks, in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston.

Marx Z., Dagan I., 2002, Conceptual Mapping through Keyword Coupled Clustering, in "Mind and Society – Special Issue on Commonsense and Scientific Reasoning".

Miller G., Charles W., 1991, Contextual Correlates of Semantic Similarity, in "Language and Cognitive Processes" 6.1: -28.

Pedersen T., Purandare A., Kulkarni A., 2005, *Name Discrimination by Clustering Similar Contexts,* in *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Schütze H., 1998, Automatic Word Sense Discrimination, in "Computational Linguistics" 24.1: 97-123.