A Comparison of Bag of Words and Syntax-based Approaches for Word Categorization

Tim Van de Cruys
Humanities Computing
University of Groningen
t.van.de.cruys@rug.nl

Abstract

This paper will examine the aptness of various word space models for the task of word categorization, as defined by the lexical semantics workshop at ESSLLI 2008. Three word clustering tasks will be examined: concrete noun categorization, concrete/abstract noun discrimination, and verb categorization. The main focus will be on the difference between bag of words models and syntax-based models. Both approaches will be evaluated with regard to the three tasks, and differences between the clustering solutions will be pointed out.

1 Introduction

For quite some years now, word space models are a popular tool for the automatic acquisition of semantics from text. In word space models, a particular word is defined by the context surrounding it. By defining a particular word (i.e. its context features) in a vector space, the word can be compared to other words, and similarity can be calculated.

With regard to the context used, two basic approaches exist. One approach makes use of 'bag of words' co-occurrence data; in this approach, a certain window around a word is used for gathering co-occurrence information. The window may either be a fixed number of words, or the paragraph or document that a word appears in. Thus, words are considered similar if they appear in similar windows (documents). One of the dominant methods using this method is LATENT SEMANTIC ANALYSIS (LSA).

The second approach uses a more fine grained distributional model, focusing on the syntactic relations that words appear with. Typically, a large text corpus is parsed, and dependency triples are extracted.[1] Words are considered similar if they appear with similar syntactic relations. Note that the former approach does not need any kind of linguistic annotation, whereas for the latter, some form of syntactic annotation is needed.

The results yielded by both approaches are typically quite different in nature: the former approach typically puts its finger on a broad, thematic kind of similarity, while the latter approach typically grasps a tighter, synonym-like similarity. Example (1) shows the difference between both approaches; for each approach, the top ten most similar nouns to the noun *muziek* 'music' are given. In (a), the window-based approach is used, while (b) uses the syntax-based approach. (a) shows indeed more thematic similarity, whereas (b) shows tighter similarity.

(1) a. **muziek** 'music': *gitaar* 'guitar', *jazz* 'jazz', *cd* 'cd', *rock* 'rock', *bas* 'bass', *song* 'song', *muzikant* 'musician', *musicus* 'musician', *drum* 'drum', *slagwerker* 'drummer'

b. **muziek** 'music': *dans* 'dance', *kunst* 'art', *klank* 'sound', *liedje* 'song', *geluid* 'sound', *poëzie* 'poetry', *literatuur* 'literature', *popmuziek* 'pop music', *lied* 'song', *melodie* 'melody'

This paper will provide results for the categorization tasks that have been defined for the lexical semantics workshop 'Bridging the gap between

[1] e.g. dependency relations that qualify *apple* might be 'object of *eat*' and 'adjective *red*'. This gives us dependency triples like $< apple, obj, eat >$.

semantic theory and computational simulations' at ESSLLI 2008.[2] The workshop provides three different categorization (clustering) tasks:

- concrete noun categorization
- abstract/concrete noun discrimination:
- verb categorization

The three tasks will be carried out according to the two approaches described above. In the evaluation of the various tasks, we will try to determine whether the difference between the 'bag of words' approach and the syntactic approach is responsible for different clustering outputs.

2 Methodology

2.1 General remarks

The research has been carried out for Dutch, mainly because this enabled us to use the Alpino parser (van Noord, 2006), a dependency parser for Dutch. The test sets that were provided have been translated into Dutch by three translators, and – when multiple translations were found – the majority translation has been taken as the final one. The frequencies of the Dutch words are by and large comparable to the frequencies of their English counterparts. High frequent words (*dog*, *cat*) and low-frequent ones (*snail*, *turtle*) in one language generally have the same order of magnitude in the other, although exceptions occur (*eagle*). Table 1 compares the frequencies of words of the *animal* class in the British National Corpus and the Twente Nieuws Corpus. The results for the other words are similar.

All data has been extracted from the TWENTE NIEUWS CORPUS (Ordelman, 2002), a corpus of $\pm$ 500M words of Dutch newspaper text. The corpus is consistently divided into paragraphs, constituting our window for the bag of words approach. The whole corpus has been parsed with the Alpino parser, and dependency triples have been extracted.

The clustering solutions have been computed with the clustering program CLUTO (Karypis, 2003), using the 'rbr' option as clustering algorithm (this is an algorithm that repeatedly bisects the matrix until the desired numbers of clusters is reached; application of this algorithm was prescribed by the workshop task description).

2.2 Bag of words

For the bag of words approach, matrices have been constructed that contain co-occurrence frequencies of nouns (verbs) together with the most frequent words of the corpus in a context window. As a context window, we selected the paragraphs of the newspaper. The resulting matrix has been adapted with POINTWISE MUTUAL INFORMATION (PMI) (Church and Hanks, 1990).

The final test matrix has been constructed in two different ways:

1. a small matrix is extracted, containing only the frequencies of the words in the test set. The output is a matrix of e.g. 45 nouns by 2K co-occurring words;
2. a large matrix is extracted, containing the frequencies of a large number of words (including the test words). The output is a matrix of e.g. 10K nouns by 2K co-occurring words. After applying PMI, the test words are extracted from the large matrix.

The choice of method has a considerable impact on the final matrix, as the results of the PMI computation are rather different. In the first case, only the test words are taken into account to normalize the features; in the second case, the features are normalized with regard to a large set of words in the corpus. The difference will lead to different clustering results. The first method will be coined LOCAL PMI (LOCPMI), the second GLOBAL PMI (GLOPMI).

We have experimented with two kinds of dimensionality reduction: LATENT SEMANTIC ANALYSIS (LSA, Landauer et al. (1997; 1998)), in which a SINGULAR VALUE DECOMPOSITION (SVD) is computed of the original co-occurrence frequency matrix[3], and NON-NEGATIVE MATRIX FACTORIZATION (Lee and Seung, 2000), in which a factorization of the original frequency is calculated by minimizing Kullback-Leibler divergence between the

[2] http://wordspace.collocations.de/doku.php/esslli:start

[3] The original method of LSA uses the frequency of words by documents as input; we used frequencies of words by co-occurring words in a context window.

NOUN.ENG	FREQ.BNC	LOGFREQ.BNC	NOUN.DU	FREQ.TWNC	LOGFREQ.TWNC
chicken	2579	7.86	kip	7663	8.94
eagle	1793	7.49	arend	113	4.72
duck	2094	7.65	eend	3245	8.08
swan	1431	7.27	zwaan	1092	7.00
owl	1648	7.41	uil	559	6.33
penguin	600	6.40	pinguïn	146	4.98
peacock	578	6.36	pauw	221	5.40
dog	12536	9.44	hond	17651	9.77
elephant	1508	7.32	olifant	2708	7.90
cow	2611	7.87	koe	9976	9.21
cat	5540	8.62	kat	5822	8.67
lion	2155	7.68	leeuw	2055	7.63
pig	2508	7.83	varken	5817	8.67
snail	543	6.30	slak	712	6.56
turtle	447	6.10	schildpad	498	6.21

Table 1: The frequencies of English words in the BNC vs. the frequencies of Dutch words in the TWNC

original matrix and its factorization according to the constraint that all values have to be non-negative. But since the dimensionality reduction models did not bring about any improvement over the simple bag of word models, dimensionality reduction models have not been included in the evaluation.

2.3 Syntax-based

The syntax-based approach makes use of matrices that contain the co-occurrence frequencies of nouns (verbs) by their dependencies. Typically, the feature space with the syntax-based method is much larger than with simple co-occurrences, but also much sparser. The resulting matrix is again adapted with PMI.

Again, the matrix can be constructed in two different ways:

1. a small matrix, containing only the frequencies of the test words by the dependencies with which the word occurs. The output is a matrix of e.g. 45 nouns by 100K dependencies;

2. a large matrix, containing the frequencies of a large number of words (including the test words). The output is e.g. a matrix of 10K nouns by 100K dependencies. The final test words are extracted afterwards.

The choice of method again has a large impact on the final matrix with regard to PMI.

2.4 Evaluation measures

There are two external evaluation measures available in CLUTO – ENTROPY and PURITY – which have been chosen as evaluation measures for the workshop task. Entropy measures how the various semantic classes are distributed within each cluster, and purity measures the extent to which each cluster contains words from primarily one class (Zhao and Karypis, 2001). Both measures run from 0 to 1. Low entropy measures and high purity values indicate a successful clustering.

3 Results & Evaluation

3.1 Concrete noun categorization

3.1.1 Introduction

In the concrete noun categorization task, the goal is to cluster 44 concrete nouns in a number of classes on various levels of generality:

- 2-way clustering: cluster nouns in two top classes *natural* and *artefact*;

- 3-way clustering: cluster nouns in three classes *animal*, *vegetable* and *artefact*;

- 6-way clustering: cluster nouns in six classes *bird*, *groundAnimal*, *fruitTree*, *green*, *tool* and *vehicle*.

In the next sections, we will evaluate how bag of words models and syntactic models are coping with this clustering task, and compare both methods.

3.1.2 Bag of words

Table 2 gives the clustering results of the bag of words methods for different clustering sizes.

method	n-way	entropy	purity
LOCPMI	2	.930	.614
	3	.489	.750
	6	.339	.636
GLOPMI	2	.983	.545
	3	.539	.705
	6	.334	.682

Table 2: A comparison of different clustering results for concrete noun categorization — bag of words approach

None of the bag of words models is particularly good at noun categorization: the LOCPMI and GLOPMI have similar results. The results do show that bag of word models are better in categorizing on a more specific level: the more specific the clustering, the better the scores are.

Figure 1 shows the confusion matrix for the GLOPMI 6-way clustering.

cluster	bird	grou	frui	gree	tool	vehi
1	0	0	0	0	1	2
2	1	0	4	5	2	0
3	0	0	0	0	0	5
4	0	0	0	0	3	0
5	6	8	0	0	0	0
6	0	0	0	0	7	0

Figure 1: Confusion matrix

The clusters found by the algorithm are still quite sensible; cluster 1 for example looks like this:

- *aardappel* ‘potatoe’, *ananas* ‘pineapple’, *banaan* ‘banana’, *champignon* ‘mushroom’, *kers* ‘cherry’, *kip* ‘chicken’, *kom* ‘bowl’, *lepel* ‘spoon’, *maïs* ‘corn’, *peer* ‘pear’, *sla* ‘lettuce’ *ui* ‘oignon’

Clearly, the algorithm has found a food-related cluster, with fruits, vegetables, a meat term (‘chicken’) and kitchen tools (‘bowl’, ‘spoon’).

The two- and three-way clusterings of the bag of words models are less sensible.

3.1.3 Syntax-based

Table 3 gives the clustering results for the syntax-based algorithms for different clustering sizes.

method	n-way	entropy	purity
LOCPMI	2	.939	.636
	3	.344	.818
	6	.118	.886
GLOPMI	2	.000	1.000
	3	.000	1.000
	6	.173	.841

Table 3: A comparison of different clustering results for concrete noun categorization — syntactic approach

LOCPMI scores the best result with regard to the most specific (6-way) clustering, but only slightly better than GLOPMI. When the clustering task becomes more abstract, GLOPMI clearly outperforms the local model: the 2-way and 3-way clusterings are optimal in the global model, whereas the local models score worse results with increasing abstractness.

Figure 2 shows the confusion matrix for the best-performing 6-way clustering (LOCPMI). The results of the global model are quite similar.

cluster	bird	grou	frui	gree	tool	vehi
1	0	0	4	5	0	0
2	0	0	0	0	7	0
3	6	0	0	0	0	0
4	0	0	0	0	0	7
5	0	0	0	0	6	0
6	1	8	0	0	0	0

Figure 2: Confusion matrix

Upon examining the results, the decisions made by the algorithm look quite reasonable:

- One bird ('chicken') is classified as *grounAnimal*;
- fruits and vegetables are assigned to one single cluster;
- the *tools* class is split up into two different clusters, so that a division is made between 'active' and 'passive' tools:
  - *beitel* 'chisel', *hamer* 'hammer', *mes* 'knife', *pen* 'pen', *potlood* 'pencil', *schaar* 'scissors', *schroevendraaier* 'screwdriver';
  - *beker* 'cup', *fles* 'bottle', *ketel* 'kettle', *kom*, 'bowl', *lepel* 'spoon', *telefoon* 'telephone'.

It is interesting to note that the difference between fruit and vegetables nonetheless is present in the data. When clustering the words from the subsets *fruitTree* and *green* into two classes, they are properly split up:

- *kers* 'cherry', *banaan* 'banana', *peer* 'pear', *ananas* 'pineapple';
- *champignon* 'mushroom', *maïs* 'corn', *sla* 'lettuce', *aardappel* 'potatoe', ui 'oignon'.

3.1.4 Comparison of both approaches

Globally, the syntax-based approach seems more apt for concrete noun clustering. Both approaches have similar results for the most specific classification (6-way clustering), but the syntax-based approach performs a lot better on a more abstract level. The conclusion might be that the bag of words approach is able to cluster nouns into 'topics' (cfr. the cluster containing words that relate to the topic 'food'), but has difficulties generalizing beyond these topics. The syntax-based approach, on the other hand, is able to generalize beyond the topics, discovering features such as 'agentness' and 'naturalness', allowing the words to be clustered in more general, top-level categories.

3.2 Abstract/Concrete Noun Discrimination

3.2.1 Introduction

The evaluation of algorithms discriminating between abstract and concrete nouns consists of three parts:

- In the first part, 30 nouns (15 with high concreteness value and 15 with low concreteness value) are clustered in two clusters, HI and LO;
- in the second part, 10 nouns with average concreteness value are added to the two-way clustering, to see whether they end up in the HI or the LO cluster;
- in the third part, a three-way clustering of the 40 nouns (15 HI, 10 ME, 15 LO) is performed.

In the next sections, both bag of word models and syntax-based models are again evaluated with regard to these parts.

3.2.2 Bag of words

Table 4 gives the clustering results of the bag of words methods for different clustering sizes.

method	part	entropy	purity
LOCPMI	part 1	.470	.867
	part 3	.505	.750
GLOPMI	part 1	.000	1.000
	part 3	.605	.700

Table 4: A comparison of different clustering results for abstract/concrete noun discrimination — bag of words approach

The GLOPMI model outperforms the LOCPMI in the discrimination of abstract and concrete nouns (part 1). The LOCPMI scores a bit better in discriminating the ME nouns (part 3).

Interestingly enough, the result of part 2 is for both LOCPMI and GLOPMI the same:

- *geur* 'smell', *vervuiling* 'pollution' and *weer* 'weather' are assigned to the HI cluster;
- *uitnodiging* 'invitation', *vorm* 'shape', *rijk* 'empire', *fundament* 'foundation', *ruzie* 'fight', *pijn* 'ache' and *ceremonie* 'ceremony' are assigned to the LO cluster.

3.2.3 Syntax-based

Table 5 gives the clustering results of the syntax-based methods for different clustering sizes.

The local PMI method gets the best results: The 2-way clustering as well as the 3-way clustering are accurately carried out.

method	part	entropy	purity
LOCPMI	part 1	.000	1.000
	part 3	.000	1.000
GLOPMI	part 1	.000	1.000
	part 3	.367	.750

Table 5: A comparison of different clustering results for abstract/concrete noun discrimination — syntactic approach

Part 2 gives different results for the LOCPMI and GLOPMI method:

- The local method classifies *rijk* 'empire' as HI and the other 9 words as LO;
- the global method classifies *weer* 'weather', *uitnodiging* 'invitation', *ceremonie* 'ceremony' as HI and the other 7 words as LO.

3.2.4 Comparison of both approaches

The syntax-based approach outperforms the bag of words approach, although the bag of words approach is also able to make an accurate distinction between concrete and abstract nouns with the GLOPMI model. Again, the explanation might be that the syntax-based method is able to discover general features of the nouns more easily. Nevertheless, the results show that discrimination between concrete and abstract nouns is possible with the bag of words approach as well as the syntax-based approach.

3.3 Verb categorization

3.3.1 Introduction

The goal of the verb categorization task is to cluster 45 verbs into a number of classes, both on a more general and a more specific level:

- 5-way clustering: cluster the verbs into 5 more general verb classes: *cognition*, *motion*, *body*, *exchange* and *changeState*;
- 9-way clustering: cluster the verbs into 9 fine-grained verb classes: *communication*, *mentalState*, *motionManner*, *motionDirection*, *changeLocation*, *bodySense*, *bodyAction*, *exchange* and *changeState*.

3.3.2 Bag of words

Table 6 gives the clustering results of the bag of words methods for different clustering sizes.

method	n-way	entropy	purity
LOCPMI	5	.478	.622
	9	.419	.578
GLOPMI	5	.463	.600
	9	.442	.556

Table 6: A comparison of different clustering results for verb categorization — bag of words approach

There are no large differences between the local and global PMI method: both methods score about the same. The more specific classification (9-way clustering) scores slightly better, but the differences are small.

Figure 3 shows the confusion matrix for the best-performing 5-way clustering (GLOPMI). The results of the local model are again similar.

cluster	cogn	moti	body	exch	chan
1	0	0	1	5	1
2	5	1	4	0	0
3	5	2	0	0	0
4	0	7	5	0	0
5	0	5	0	0	4

Figure 3: Confusion matrix

The first cluster mainly contains exchange verbs. The second cluster is a combination of cognition and body verbs. It is interesting to note that the body verbs with a particular emotional connotation ('cry', 'smile', also 'listen' and 'feel') end up in a cluster together with the cognition verbs. The body verbs without an emotional connotation ('breathe', 'drink', 'eat', also 'smell') end up in a cluster together with (body) movements (cluster 4). Cluster 3 seems a business-related cluster, given the fact that *bestuur* is an ambiguous verb in Dutch, meaning 'to drive' as well as 'to manage'.

The complete clustering is given below:

- *betaal* 'pay', *herstel* 'repair', *koop* 'buy', *leen* 'lend' , *merk* 'notice', *verkoop* 'sell', *verwerf* 'acquire'

- *ga_weg* ‘leave’, *herinner* ‘remember’, *huil* ‘cry’, *lach* ‘smile’, *lees* ‘read’, *luister* ‘listen’, *praat* ‘talk’, *vergeet* ‘forget’, *voel* ‘feel’, *weet* ‘know’

- *bestuur* ‘drive’, *controleer* ‘check’, *evalueer* ‘evaluate’, *spreek* ‘speak’, *suggereer* ‘suggest’,*verzoek* ‘request’, *zend* ‘send’

- *adem* ‘breathe’, *beweeg* ‘move’, *draag* ‘carry’, *drink* ‘drink’, *duw* ‘push’, *eet* ‘eat’, *kijk* ‘look’, *loop* ‘run’, *ruik* ‘smell’, *sta_op* ‘rise’, *trek* ‘pull’, *wandel* ‘walk’

- *breek* ‘break’ , *dood* ‘kill’ , *ga_binnen* ‘enter’, *kom_aan* ‘arrive’, *rij* ‘ride’, *sterf* ‘die’, *val* ‘fall’, *verniel* ‘destroy’, *vlieg* ‘fly’

3.3.3 Syntax-based

Table 7 gives the clustering results of the syntax-based methods for different clustering sizes.

method	n-way	entropy	purity
LOCPMI	5	.516	.644
	9	.432	.489
GLOPMI	5	.464	.667
	9	.408	.556

Table 7: A comparison of different clustering results for verb categorization — syntactic approach

The global PMI approach yields slightly better results than the local one, but differences are again small. The more specific clustering is slightly better than the more general one.

Figure 4 shows the confusion matrix for the best-performing 5-way clustering (GLOPMI).

cluster	cogn	moti	body	exch	chan
1	0	8	1	0	0
2	1	2	0	5	1
3	7	0	3	0	0
4	2	2	0	0	4
5	0	3	6	0	0

Figure 4: Confusion matrix

The first cluster contains many motion verbs; the second one has many exchange verbs, and the third one contains many cognition verbs. The fourth cluster contains mainly change verbs, but also non-related cognition and motion verbs, and the fifth one contains mostly motion verbs.

The complete clustering is given below:

- *beweeg* ‘move’, *duw* ‘push’, *kijk* ‘look’, *loop* ‘run’, *rij* ‘ride’, *trek* ‘pull’, *vlieg* ‘fly’, *wandel* ‘walk’, *zend* ‘send’

- *bestuur* ‘drive’, *betaal* ‘pay’, *controleer* ‘check’, *draag* ‘carry’, *koop* ‘buy’, *leen* ‘lend’, *verkoop* ‘sell’, *verniel* ‘destroy’, *verwerf* ‘acquire’

- *adem* ‘breathe’, *herinner* ‘remember’, *lees* ‘read’, *merk* ‘notice’, *praat* ‘talk’, *spreek* ‘speak’, *sugerreer* ‘suggest’, *vergeet* ‘forget’, *voel* ‘feel’, *weet* ‘know’

- *breek* ‘break’, *dood* ‘kill’, *evalueer* ‘evaluate’, *herstel* ‘repair’, *kom_aan* ‘arrive’, *sterf* ‘die’, *val* ‘fall’, *verzoek* ‘request’

- *drink* ‘drink’, *eet* ‘eat’, *ga_binnen* ‘enter’, *ga_weg* ‘leave’, *huil* ‘cry’, *lach* ‘smile’, *luister* ‘listen’, *ruik* ‘smell’, *sta_op* ‘rise’

3.3.4 Comparison of both approaches

The performance of the bag of words model and the syntax-based model is similar; neither of both really outperforms the other. The more specific clustering solutions are slightly better than the general ones.

There is considerable difference between the clustering solutions found by the bag of words approach and the syntax-based approach. Again, this might be due to the kind of similarity found by the models. The bag of words approach seems to be influenced by topics again (the business related cluster), whereas the syntax-based model might be influenced by more general features (‘motion’ in the first cluster). But given the evaluation results of the verb clustering, these are very tentative conclusions.

4 Conclusions & Future Work

The evaluation results presented in the former section indicate that semantic space models are fruitful models for the induction of actual semantic classes.

Especially the noun categorizations – the concrete noun categorization and the concrete/abstract noun discrimination task – perform very well. Verb categorization is a more difficult task for the semantic models presented in this paper: the results are worse than those for the noun categorizations.

In general, the syntax-based approach yields better results than the bag of words approach. This might be due to the fact that bag of words models get at a kind of topical semantic similarity, whereas the syntax-based model might be able to extract more abstract properties of the words. These are, however, tentative conclusions for a tendency present in the data. More research is needed to found this statement.

In most cases, a global method of calculation PMI yields better results. If the algorithm is able to normalize the word vectors according to the distributions of a large number of words in the data, the clustering solution is generally better. There are, however, some exceptions.

In the future, other semantic models need to be investigated for the categorization of verbs. Including subcategorization information in the models might be beneficial for the clustering of verbs, as well as generalizing among the feature space of the verb's dependencies (e.g. by using semantic noun clusters instead of nouns).

One last issue for future work is the comparison between small clustering tasks like the ones presented above, and the clustering solutions of a large clustering framework, in which a large number of words are captured.

Nevertheless, the results of the present evaluation tasks indicate that word space models are a suitable tool for the induction of semantic classes.

References

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

Z. Harris. 1985. Distributional structure. In Jerrold J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press.

George Karypis. 2003. CLUTO - a clustering toolkit. Technical Report #02-017, nov.

Thomas Landauer and Se Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychology Review*, 104:211–240.

Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:295–284.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562.

R.J.F. Ordelman. 2002. Twente Nieuws Corpus (TwNC), August. Parlevink Language Techonology Group. University of Twente.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Leuven.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical Report #01-40.