Factored Translation with Unsupervised Word Clusters

Christian Rishøj
Center for Language Technology
University of Copenhagen
crjensen@hum.ku.dk

Anders Søgaard
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Abstract

Unsupervised word clustering algorithms — which form word clusters based on a measure of distributional similarity — have proven to be useful in providing beneficial features for various natural language processing tasks involving supervised learning. This work explores the utility of such word clusters as factors in statistical machine translation.

Although some of the language pairs in this work clearly benefit from the factor augmentation, there is no consistent improvement in translation accuracy across the board. For all language pairs, the word clusters clearly improve translation for some proportion of the sentences in the test set, but has a weak or even detrimental effect on the rest.

It is shown that if one could determine whether or not to use a factor when translating a given sentence, rather substantial improvements in precision could be achieved for all of the language pairs evaluated. While such an "oracle" method is not identified, evaluations indicate that unsupervised word cluster are most beneficial in sentences *without* unknown words.

1 Factored translation

One can go far in terms of translation quality with plenty of bilingual text and a translation model that maps small chunks of tokens as they appear in the surface form, that is, the usual phrase-based statistical machine translation model. Yet even with a large parallel corpus, data sparsity is still an issue. Factored translation models are an extension of phrase-based models which allow integration of additional word-level annotation into the model. Operating on more general representations, such as lemmas or some kind of stems, translation model can draw on richer statistics and to some degree offset the data sparsity problem.

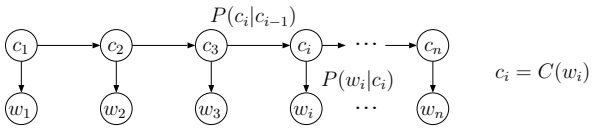

Figure 1: Bayesian network illustrating the class-based language model that is used to define the quality of a clustering in the Brown algorithm [Liang, 2005]

2 Unsupervised word clusters

Unsupervised word clusters owe their appeal perhaps mostly to the relative ease of obtaining them. Obtaining regular morphological, syntactic or semantic analyses for tokens in a text relies on some sort of tagger, either based on manually crafted rules or trainable on an annotated corpus. Both rule-crafting and corpus annotation are time-consuming and expensive processes, and might not be feasible for a small or resource-scarce language.

For unsupervised word clusters, on the other hand, one merely needs a large amount of raw (unannotated) text and some processing power. Such clustering is thus particularly interesting for resource-scarce languages, and especially so if the clusters enable the training of more generalized translation models without more bilingual text.

The independence of annotated corpora or handcrafted rules make unsupervised clusters interesting for languages rich in NLP resources too. They offer a way to exploit vast amounts of raw, unannotated, monolingual text, in a manner akin to the way language models profitably may be trained on vast amounts of raw monolingual text.

With the broad coverage achievable from vast amounts of monolingual text, word clusters might help alleviate the problem of unknown words in translation. It is imaginable that a word form otherwise unknown to the translation model belongs to

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 447–451,
Edinburgh, Scotland, UK, July 30–31, 2011.

a known cluster. Appropriate use of word clusters, coupled with a broad-coverage language model, could make it be possible for the translation model to arrive at the intended translation.

In this work we use two unsupervised clustering algorithms: Brown and Unsupos. Other clustering algorithms were on the drawing board as well, namely embeddings from the Neural Language Model of Collobert and Weston [2008] and word representations from random indexing (RI)[1]. These, however, were abandoned due to time constraints.

2.1 The Brown algorithm

The bottom-up agglomerative algorithm of Brown et al. [1992] processes a sequence of tokens and produces a binary tree with tokens as leaf nodes. Each internal node in the tree can be interpreted as a cluster containing the tokens on the leaf nodes of that subtree. The clustering produced is thus a *hierarchical* clustering.

Very briefly, the algorithm proceeds by first assigning every token to its own cluster, and then iteratively merges the two clusters that maximises the quality of the resulting clustering, where the *quality* of a clustering is defined in terms of a *class-based language model* (figure 1).

Note that this algorithm produces a *hard clustering*, in the sense that it assigns each token to a single cluster. From a semantic perspective, there are homographic words whose underlying senses are conceptually and possibly syntactically distinct, and whose cluster-tag intuitively should depend on their use in running text. The clustering obtained from the Brown algorithm does not accommodate this wish.

We use the implementation[2] of Liang [2005].

2.2 jUnsupos

Contrary to the hard clustering of the Brown algorithm, the jUnsupos algorithm of Biemann [2006] emits a Viterbi tagger which is sensitive to the context of a token in running text. Thus, word forms can belong to more than a single cluster, and such word forms — which are considered ambiguous by the algorithm — will be assigned to a cluster depending on their context.

In a coarse outline, the algorithm works by first inducing a distributional clustering for unambiguous high-frequency tokens, as well as a co-occurrence-based clustering for less common tokens. The two partly overlapping clusterings are then combined to produce a lexicon with derived syntactic categories and word forms.

[1] https://github.com/turian/random-indexing-wordrepresentations

[2] Available at http://www.cs.berkeley.edu/~pliang/software/

100001001 immediate urgent ongoing absolute extraordinary exceptional ideological unprecedented appalling overwhelming alleged automatic [...]

11111100111111110 worried concerned skeptical unhappy uneasy reticent unsure perplexed excited apprehensive legion unconcerned [...]

111111100010001 cover include involve exclude confuse encompass designate preclude transcend duplicate defy precede [...]

1111111000000 encourage promote protect defend safeguard restore assist preserve coordinate convince destroy integrate [...]

0111000 china russia iran israel turkey ukraine india japan pakistan georgia serbia europol [...]

1000110010 waste water drugs land fish material meat profit alcohol forest blood chemicals [...]

Figure 2: Exemplars of word clusters obtained using the Brown algorithm (C=1000), showing the 12 most frequent tokens per cluster

2.3 Cluster count and complexion

A reasonable question when faced with the task of inducing word clusters in an unsupervised manner is: How many clusters to produce? This question is presumably closely intertwined with the question of what sort of beast a cluster obtained in this manner can be expected to be. Would a clustering with around 30-90 clusters correspond somewhat closely to an ordinary part-of-speech tag-set for the given language?

Looking at the handful of exemplar clusters shown in figure 2, which were obtained with the Brown algorithm (using a cluster count of 1000), we cautiously note some apparent patterns.

- The clusters appear to be subsets of the clustering implied by conventional part-of-speech tags: The first two consist of adjectives (including the rather ambiguous form *legion*), the next two (transitive) verbs and the final two nouns.

- Syntactically, members of the two apparent verb

clusters seem to consist of verbs in their infinitive (or plurally inflected) form.

- From a quasi-semantic perspective, the last cluster appears to consist of nouns for corporeal goods (as apposed to immaterial things).
- While most exemplars from the second-last cluster are countries, all of the shown forms can be said to be proper nouns.

Note that only the 12 most frequent forms from each cluster are displayed, the apparent patterns should be taken with a pinch of salt. Although the qualities suggested can be expected to relate to distributional properties that the clusters reflect, exceptional members are perhaps to be expected.

In the present work, we went with the pre-trained models for jUnsupos[3], which have the following characteristics[4]:

Lang	Corpus	# Sents	# Tags
cs	LCC	4 M	539
de	Wortschatz	40 M	396
en	Medline 2004	34 M	480
es	LCC	4.5 M	415
fr	LCC	3 M	359

For the Brown algorithm, we are contrasting cluster count choices of 320 and 1000, based on reports of other successful applications [Turian et al., 2010][5], with clustering models trained on monolingual data from the Europarl corpus and the News Commentary corpus.

3 Experimental setup

The baseline systems were set up in accordance with the guidelines on the shared task website. That is, they were trained with `grow-diag-final-and` word alignment heuristics and `msd-bidirectional-fe` reordering.

Translation models were trained on a concatenation of the Europarl and News Commentary corpora, which were first tokenized, then filtered to sentence lengths of up to 40 tokens, and finally lowercased.

5-gram language models were built using `ngram-count` on a concatenation of the Europarl corpora and the News Commentary corpora. For the unsupervised word clusters, 5-gram language models were used as well, built from tagged versions of the same corpora. All language models were binarised and loaded using KenLM [Heafield, 2011].

Minimum error rate training (MERT) was used to optimise parameters on both baseline and factored models against the 2008 news test set, as suggested on the shared task website[6].

All phrase tables were filtered and binarised for the development and testing corpora during tuning and testing, respectively.

Seeing that the preparation of the raw corpora, word clustering models, factored corpora, language models, as well as training, optimization and evaluation of the various models was a rather involved, yet repetitive process, we took a stab at making a GNU Makefile-based approach for automated handling (and parallelisation) of the whole dependency graph of subtasks. The ongoing effort, which shares some aspirations and abilities with the recently announced Experiment Management System (EMS), is publicly available[7].

4 Results

Table 1a lists BLEU scores for adding jUnsupos tags (*uPOS*), Brown clusters with 320 clusters (*C320*) or Brown clusters with 1000 clusters (*C1000*) as either an alignment factor, a two-sided translation factor or a source-sided translation factor.

Although using Brown clusters (C1000) as a two-sided translation factor improves BLEU scores for some language pairs, most notably *en-cs*, *en-de* and *cs-en*, no clear across-the-board benefit is seen.

4.1 Oracle scores

Based on the hypothesis that the factorisations are beneficial when translation some sentences, and not when translating others, we completed an oracle-based evaluation, in which we assume to know *a priori* whether to use the factored model for translating a given sentence, or just go with the baseline, unfactored model. In reality, we don't have such an oracle method for arbitrary sentences, but when dealing with the shared task test set (or other corpora for which we have reference translations), it was easy enough to check per-sentence BLEU scores for each model and make the decision based on a comparison.

Table 1b lists BLEU scores obtainable with each factor configuration given such an oracle method. In this scenario, most factored models beat the baseline, indicating that the factorisations are beneficial for certain sentences, and detrimental for others.

[3] As available at http://wortschatz.uni-leipzig.de/~cbiemann/software/unsupos.html

[4] LCC refers to the Leipzig Corpora, available at http://corpora.uni-leipzig.de/. Wortschatz refers to http://www.wortschatz.uni-leipzig.de/. Medline is available at http://www.nlm.nih.gov/mesh/filelist.html.

[5] A planned evaluation of a cluster count of 3200 was abandoned due to time constraints

[6] http://www.statmt.org/wmt11/translation-task.html

[7] At https://gibhub.com/crishoj/factored

Pair	Baseline	Alignment factor			Two-sided translation			Source-sided transl.			Best	
		C1000	C320	uPOS	C1000	C320	uPOS	C1000	C320	uPOS	Δ	%
cs-en	18.18	17.77	17.19	13.54	**18.59**	18.36	17.50	18.19	18.19	17.59	*0.41*	2.3%
de-en	18.45	17.94	17.57	16.36	**18.56**	18.42	17.93	18.12	18.12	17.86	*0.11*	0.6%
en-cs	11.85	11.82	11.61	9.75	**12.73**	12.28	10.94	11.92	11.92	11.85	*0.88*	7.4%
en-de	13.27	12.90	12.83	11.98	13.81	**13.84**	13.19	12.94	12.94	12.92	*0.57*	4.3%
en-es	28.08	27.10	26.52	24.90	**28.40**	28.16	27.50	27.31	27.31	27.19	*0.32*	1.1%
en-fr	**25.90**	24.60	23.98	21.85	25.89	20.59	24.16	24.89	24.89	24.74	–	–
es-en	**26.70**	24.87	24.71	23.92	25.76	25.96	25.40	24.92	24.92	24.92	–	–
fr-en	**24.73**	23.18	23.13	21.76	24.01	22.86	23.23	23.37	23.37	23.04	–	–

(a) BLEU scores for factor configurations in comparison to the unfactored baseline

Pair	Baseline	Alignment factor			Two-sided translation			Source-sided transl.			Best	
		C1000	C320	uPOS	C1000	C320	uPOS	C1000	C320	uPOS	Δ	%
cs-en	18.18	19.93	19.81	19.19	**20.01**	20.00	19.83	19.58	19.58	19.63	1.83	10.1%
de-en	18.45	20.06	20.00	19.75	**20.28**	20.26	20.15	19.84	19.84	19.90	1.83	9.9%
en-cs	11.85	13.18	13.14	12.81	**13.77**	13.58	12.98	12.83	12.83	12.93	1.92	16.2%
en-de	13.27	14.56	14.60	14.36	14.98	**15.10**	14.81	14.21	14.21	14.28	1.83	13.8%
en-es	28.08	29.70	29.50	29.17	**30.33**	30.2	30.00	29.54	29.54	29.56	2.25	8.0%
en-fr	25.90	27.34	27.22	26.90	**27.84**	26.98	27.32	27.15	27.15	27.16	1.94	7.5%
es-en	26.70	27.83	27.81	27.74	28.16	**28.20**	28.06	27.64	27.64	27.73	1.50	5.6%
fr-en	24.73	25.86	25.95	25.83	26.16	**26.31**	26.05	25.66	25.66	25.69	1.58	6.4%

(b) BLEU scores with an *oracle*-directed, per-sentence selective usage of either the baseline or the factored model

Table 1: BLEU scores when using Brown Clusters with granularity 1000 (*C1000*), granularity 320 (*C320*) and unsupervised part-of-speech tags (*uPOS*) as either an added alignment factor, a two-sided translation factor or a source-sided translation factor

Pair	Baseline	Oracle	Abs. Δ	Rel. %
cs-en	18.18	22.60	4.42	24.3%
de-en	18.45	22.42	3.97	21.5%
en-cs	11.85	15.89	4.04	34.1%
en-de	13.27	17.16	3.89	29.3%
en-es	28.08	32.52	4.44	15.8%
en-fr	25.90	30.07	4.17	16.1%
es-en	26.70	30.22	3.52	13.2%
fr-en	24.73	28.67	3.94	15.9%

Table 2: BLEU scores under the assumption of an oracle function indicating the optimal factor configuration for each sentence

4.2 Combined oracle scores

Imagine another oracle function, which would not simply determine whether to prefer a given factored model over the baseline for a given sentence, but instead indicate which of several possible factored models to use when translating a given sentence.

BLEU scores obtainable under the assumption of such a combined oracle function are listed in table 2. As was the case for the individual factored models (table 1a), *en-cs*, *en-de* and *cs-en* see the largest benefits over the baselines.

These oracle scores are obviously an idealised case. They indicate an upper bound that one could seek to approximate by constructing an appropriate oracle function.

4.3 Unknown words

In section 2 it was hypothesised that word clusters are potentially beneficial in translating sentences with unknown words — that is, word forms which were not seen in any aligned sentences (but which may belong to a word cluster known by the translation model).

With this hypothesis in mind, we would like to

Pair	Sentences		Baseline	C1000	Rel. %
cs-en	1955	65%	17.63	**17.70**	0.4%
de-en	1925	64%	**17.84**	17.56	-1.6%
en-cs	1583	53%	11.85	**12.63**	6.6%
en-de	1395	46%	**13.65**	13.47	-1.3%
en-es	1327	44%	27.77	**27.97**	0.7%
en-fr	1369	46%	**25.43**	25.11	-1.3%
es-en	1316	44%	**26.43**	25.41	-3.9%
fr-en	1423	47%	**24.20**	23.56	-2.6%
Avg.	*1537*	*51%*	***20.60***	*20.43*	*-0.4%*

(a) BLEU scores for sentences *with* unknown words

Pair	Sentences		Baseline	C1000	Rel. %
cs-en	1048	35%	19.63	**20.77**	5.8%
de-en	1078	36%	20.03	**21.24**	6.0%
en-cs	1420	47%	11.85	**12.90**	8.9%
en-de	1608	54%	12.97	**14.22**	9.6%
en-es	1676	56%	28.41	**28.88**	1.7%
en-fr	1634	54%	26.46	**26.81**	1.3%
es-en	1687	56%	**27.01**	26.15	-3.2%
fr-en	1580	53%	**25.40**	24.58	-3.2%
Avg.	*1466*	*49%*	*21.47*	***21.94***	*3.4%*

(b) BLEU scores for sentences with *no* unknown words

Table 3: BLEU scores for the best overall factorisation, Brown clusters (C=1000) as a two-sided translation factor, on sentences *with* (table 3a) and *without* (table 3b) unknown words

see how the factored models fare in comparison to the unfactored baselines, specifically for those sentences containing unknown words, and for the rest (sentences *without* unknown words). This targeted evaluation was done using the best overall factor configuration: Brown clusters (C=1000) as a two-sided translation factor.

The results are shown in tables 3a and 3b. On average (across language paris), 51% test set sentences contain at least 1 unknown word. Contrary to what might be expected, the factorisation seems to be most beneficial for sentences with all *known* words (3.4% improvement in BLEU score on average). For sentences with unknown words, the effect is weak or detrimental (except for *en-cs*), averaging a slight decrease (-0.4%) in BLEU score across the language pairs.

The lack of benefit for sentences with unknown words is likely due to the fact that no additional monolingual data was used to make the Brown clusters for this experiment. In other words, there is no chance of knowing the Brown cluster for an unknown word. Furthermore, we assume that gains for sentences with unknown words are more likely with a factorisation that includes an alternative decoding path for word clusters[8].

5 Conclusions and future work

In this work we have explored the utility of three unsupervised word clusterings as either an alignment factor, a two-sided translation factor or a source-sided translation factor.

Although no across-the-board benefit was seen, it was evident that the factorisations help in translating some proportion of the test set sentences. Being able to determine for which sentences to use a factored model is clearly desirable.

Overall, the single most beneficial of the factor configurations explored was Brown clusters with a granularity of 1000, as a two-sided translation factor. A more detailed evaluation of the effects of different cluster sizes, as well as using clusters induced from more text, would be interesting in a follow-up study.

Using clusters in some more interesting factor configurations, particularly in alternative decoding paths, is still pending.

References

C. Biemann. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 7–12, 2006.

P. F Brown, V. J.D Pietra, P. V deSouza, J. C Lai, and R. L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, page 160–167, 2008.

K. Heafield. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK, July 2011. Association for Computational Linguistics.

P. Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.

J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 384–394, 2010.

[8]Evaluation of factor configurations with alternative decoding paths were abandoned due to limited computational resources and initially discouraging results