Sense-Aware Statistical Machine Translation using Adaptive Context-Dependent Clustering

Xiao Pu
EPFL & Idiap Research Inst.
Martigny, Switzerland
xiao.pu@idiap.ch

Nikolaos Pappas
Idiap Research Institute
Martigny, Switzerland
nikolaos.pappas@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
apbelis@idiap.ch

Abstract

Statistical machine translation (SMT) systems use local cues from n-gram translation and language models to select the translation of each source word. Such systems do not explicitly perform word sense disambiguation (WSD), although this would enable them to select translations depending on the hypothesized sense of each word. Previous attempts to constrain word translations based on the results of generic WSD systems have suffered from their limited accuracy. We demonstrate that WSD systems can be adapted to help SMT, thanks to three key achievements: (1) we consider a larger context for WSD than SMT can afford to consider; (2) we adapt the number of senses per word to the ones observed in the training data using clustering-based WSD with K-means; and (3) we initialize sense-clustering with definitions or examples extracted from WordNet. Our WSD system is competitive, and in combination with a factored SMT system improves noun and verb translation from English to Chinese, Dutch, French, German, and Spanish.

1 Introduction

Selecting the correct translation of polysemous words remains an important challenge for machine translation (MT). While some translation options may be interchangeable, substantially different senses of source words must generally be rendered by different words in the target language. In this case, an MT system should identify – implicitly or explicitly – the correct sense conveyed by each occurrence in order to select the appropriate translation.

Source: And I do really like this *shot*, because it shows all the detritus that's sort of embedded in the sole of the sneakers.

Baseline SMT: Und ich mag dieses *Bild* ...

Online NMT: Und ich mag diesen *Schuss* wirklich, ...

Sense-aware MT: Und ich mag diese *Aufnahme* wirklich, ...

Reference translation: Ich mag diese *Aufnahme* wirklich, ...

Figure 1: Example of sense-aware translation that is closer to a reference translation than a baseline statistical MT system or an online neural one.

Current statistical or neural MT systems perform word sense disambiguation (WSD) implicitly, for instance through the n-gram frequency information stored in the translation and language models. However, the context taken into account by an MT system when performing implicit WSD is limited. For instance, in the case of phrase-based SMT, it is the order of the language model (often between 3 and 5) and the length of n-grams in the phrase table (seldom above 5). In attention-based neural MT systems, the context extends to the entire sentence, but is not specifically trained to be used for WSD.

For instance, Figure 1 shows an English sentence translated into German by a baseline statistical MT, an online neural MT, and the sense-aware MT system proposed in this paper. The word *shot* is respectively translated as *Schuss* (gun shot), *Bild* (drawing) and *Aufnahme* (picture) by the online NMT, the baseline system, and our sense-aware system. The latter selects a correct sense, which is identical to the reference translation, while the first two are incorrect (especially the online NMT).

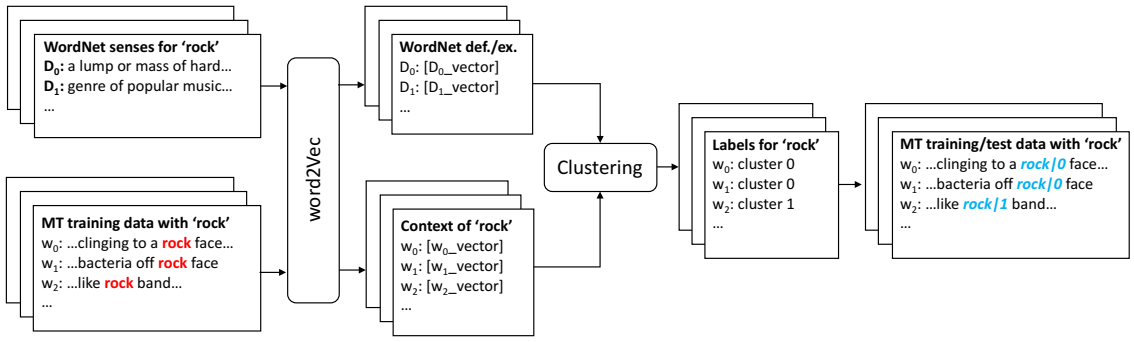

Figure 2: Adaptive WSD for MT: vectors from WordNet definitions (or examples) are clustered with context vectors of each occurrence (here of *'rock'*), resulting in sense labels used as factors for MT.

In this paper, we introduce a sense-aware statistical MT system that performs explicit WSD, and uses for this task a larger context than is accessible to state-of-the-art SMT. Our WSD system performs context-dependent clustering of word occurrences and is initialized with knowledge from WordNet, in the form of vector representations of definitions or examples for each sense. The labels of the resulting clusters are used as abstract source-side sense labels within a factored phrase-based SMT system. The stages of our method are presented in Figure 2, and will be explained in detail in Section 3.

Our results, presented in Section 5, show first that our WSD system is competitive on the SemEval 2010 WSD task, but especially that it helps SMT to increase its BLEU scores and to improve the translation of polysemous nouns and verbs, when translating from English into Chinese, German, French, Spanish or Dutch, in comparison to an SMT baseline that is not aware of word senses.

With respect to previous work that used WSD for MT, discussed in Section 2, we innovate on the following points:

- we design a sense clustering method with explicit knowledge (WordNet definitions or examples) to disambiguate polysemous nouns and verbs;
- we represent each token by its context vector, obtained from word2vec word vectors in a large window surrounding the token;
- we adapt the possible number of senses per word to the ones observed in the training data rather than constraining them by the full list of senses from WordNet;
- we use the abstract sense labels for each analyzed word as factors in an SMT system.

2 Related Work

Word sense disambiguation aims to identify the sense of a word appearing in a given context (Agirre and Edmonds, 2007). Resolving word sense ambiguities should be useful, in particular, for lexical choice in MT.

An initial investigation found that an SMT system which makes use of off-the-shelf WSD does not yield significantly better quality translations than a SMT system not using it (Carpuat and Wu, 2005). However, another study (Vickrey et al., 2005) reformulated the task of WSD for SMT as predicting possible target translations rather than senses of ambiguous source words, and showed that WSD improved such a simplified word translation task. Subsequent studies which adopted this formulation (Cabezas and Resnik, 2005; Chan et al., 2007; Carpuat and Wu, 2007), successfully integrated WSD to hierarchical or phrase-based SMT. These systems yielded slightly better translations compared to SMT baselines in most cases (0.15–0.30 BLEU).

Although the WSD reformulation above proved helpful for SMT, it did not determine whether actual source-side senses are helpful or not for end-to-end SMT. Xiong and Zhang (2014) attempted to answer this question by performing word sense induction for large scale data. In particular, they proposed a topic model that automatically learned sense clusters for words in the source language. In this way, on the one hand, they avoided using a pre-specified inventory of word senses as traditional WSD does, but on the other hand, they created the risk of discovering sense clusters which do not correspond to the common senses of words needed for MT. Hence, this study left open an important question, namely whether WSD based on

semantic resources such as WordNet (Fellbaum, 1998) can be successfully integrated with SMT.

Neale et al. (2016) attempted such an integration, by using a WSD system based on a sense graph from WordNet (Agirre and Soroa, 2009). This system detects the senses of words in context using a random walk algorithm over the sense graph. The authors used it to specify the senses of the source words and integrate them as contextual features with a MaxEnt-based translation model for English-Portuguese MT. Similarly, Su et al. (2015) built a large weighted graph model of both source and target word dependencies and integrated them as features to a SMT model. However, apart from the sense graph, WordNet provides also textual information such as sense definitions and examples, which should be useful for disambiguating senses, but were not used in the above studies. Here, we aim to exploit this information to perform word sense induction from large scale monolingual data (in a first phase), thus combining the benefits of semantic ontologies and word sense induction for WSD.

Several other studies integrated additional information from a larger context using factored-based MT models (Koehn and Hoang, 2007). Birch et al. (2007) used supertags from a Combinatorial Categorial Grammar as factors in phrase-based translation model. Avramidis and Koehn (2008) added source-side syntactic information for each word for translating from a morphologically poorer language to a richer one (English-Greek). The levels of improvement achieved with factored models such as the ones above range from 0.15 to 0.50 BLEU points. Here, we also observe improvements in the upper part of this range, and they are consistent across several language pairs.

3 Adaptive Sense Clustering for SMT

In this section, we describe our adaptive WSD method and show how we integrate it with SMT, as represented in Figure 2 above. In a nutshell, we consider all source words that have more than one sense (synset) in WordNet, and extract from WordNet the definition of each sense and, if available, the example. We associate to them word embeddings built using word2vec. For each occurrence of these words in the training data, we also build vectors for their contexts (i.e. neighboring words) using the same model. All the vectors are passed to a clustering algorithm, resulting in the labeling of each occurrence with a cluster number that will be used as a factor in statistical MT.

Our method answers several limitations of previous supervised or unsupervised WSD methods. Supervised methods require data with manually sense-annotated labels and are therefore often limited to a small number of word types: for instance, only 50 nouns and 50 verbs were targeted in SemEval 2010[1] (Manandhar et al., 2010). On the contrary, our method does not require labeled texts for training, and applies to all word types appearing with multiple senses in WordNet.

Unsupervised methods often pre-define the number of possible senses for each ambiguous word before clustering the various occurrences according to the senses. If these numbers come from WordNet, the senses may be too fine-grained for the needs of translation, especially when a specific domain is targeted. In contrast, as we explain below, our WSD method initializes a context-dependent clustering algorithm with information from WordNet senses for each word (nouns and verbs), but then adapts the number of clusters to the observed training data for MT.

3.1 Representing Definitions, Examples and Contexts of Word Occurrences

For each noun or verb *type* W_t appearing in the training data, as identified by the Stanford POS tagger,[2] we extract the senses associated to it in WordNet[3] by using NLTK.[4] Specifically, we extract the set of definitions $D_t = \{d_{tj}|j = 1, \ldots, m_t\}$ and the set of examples of use $E_t = \{e_{tj}|j = 1, \ldots, n_t\}$, each of them containing multiple words. While most of the senses are accompanied by a definition, only a smaller subset also include an example of use, as it appears from the four last columns of Table 1. Less frequently, some senses contain examples without definitions.

Each definition d_{tj} and example e_{tj} is represented by a vector, which is the average of the word embeddings over all the words constituting them (except stopwords). Formally, these are $\vec{d_{tj}} = (\sum_{w_l \in d_{tj}} \vec{w_l})/m_t$ and respectively $\vec{e_{tj}} = (\sum_{w_l \in e'_{tj}} \vec{w_l})/n_t$. While the entire definition d_{tj} is used to build the vector, we do not consider all words in the example e_{tj}, but limit the sum to

[1] www.cs.york.ac.uk/semeval2010_WSI
[2] http://nlp.stanford.edu/software/
[3] https://wordnet.princeton.edu/
[4] See www.nltk.org/howto/wordnet.html

e'_{tj}i.e. we consider only a window of size c centered around the noun or verb of type W_t (similarly to the window used for context representation below) to avoid noise from long examples.

All the word vectors $\vec{w_l}$ above are word2vec pre-trained embeddings from Google[5] (Mikolov et al., 2013). If d is the dimensionality of the word vector space, then all vectors $\vec{w_l}$, $\vec{d_{tj}}$, and $\vec{e_{tj}}$ are in $\mathcal{R}^d$. Each definition vector $\vec{d_{tj}}$ or example vector $\vec{e_{tj}}$ for a word type W_t will be considered as a center vector for each sense during the clustering procedure.

Similarly, each word *token* w_i in a source sentence is represented by the average vector $\vec{u_i}$ of the words in its context, which is defined as a window of c words centered in w_i. The value c of the context size is even, since we calculate the vector $\vec{u_i}$ for w_i by averaging vectors from $c/2$ words before w_i and from $c/2$ words after it. We stop nevertheless at the sentence boundaries, and filter out stop words before averaging.

We will now explain how to cluster according to their senses all vectors $\vec{u_i}$ for the occurrences w_i of a given word type W_t, using as initial centers either the definition or the example vectors.

3.2 Clustering Word Occurrences According to their Senses

We aim to group all occurrences w_i of a given word type W_t into clusters according to the similarity of their senses, which we will model as the similarity of their context vectors. The correctness of this hypothesis will be supported by the empirical results. We will modify the k-means algorithm in several ways to achieve an optimal clustering of word senses for MT.

The original k-means algorithm (MacQueen, 1967) aims to partition a set of items, which are here tokens $w_1, w_2, \ldots, w_n$ of a same word type W_t, represented through their embeddings $\vec{u_1}, \vec{u_2}, \ldots, \vec{u_n}$ where $\vec{u_i} \in \mathcal{R}^d$. The goal of k-means is to partition (or cluster) them into k sets $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares, as follows:

$$S = \arg\min_{S} \sum_{i=1}^{k} \sum_{\vec{u} \in S_i} ||\vec{u} - \vec{\mu_i}||^2, \qquad (1)$$

where $\vec{\mu_i}$ is the centroid of each set S_i. At the first iteration, when there are no clusters yet, the algorithm selects k random points to be the centroids of the k clusters. Then, at each subsequent iteration t, k-means calculates for each candidate cluster a new point to be the centroid of the observations, defined as their average vector, as follows:

$$\vec{\mu_i}^{t+1} = \frac{1}{|S_i^t|} \sum_{\vec{u_j} \in S_i^t} \vec{u_j} \qquad (2)$$

We make the following modifications to the original k-means algorithm, to make it adaptive to the word senses observed in the training data.

1. We define the initial number of clusters k_t for each ambiguous word type W_t in the data as the number of its senses in WordNet (but this number may be reduced by the final re-clustering described below at point 3). Specifically, we run two series of experiments (the results of which will be compared in Section 5.1.1): one in which each k_t is set to m_t, i.e. the number of senses that possess a definition in WordNet, and another one in which we consider only senses that are illustrated with an example, hence setting each k_t to n_t. These settings avoid fixing the number of clusters k_t arbitrarily for each ambiguous word type.

2. We initialize the centroids of the clusters to the vectors representing the senses from WordNet, either using their definition vectors $\vec{d_{tj}}$ in one series of experiments, or their example vectors $\vec{e_{tj}}$ in the other one. This second modification attempts to provide a reasonably accurate starting point for the clustering process.

3. After running the k-means algorithm, we reduce the number of clusters for each word type by merging the clusters which contain fewer than 10 tokens with the nearest larger cluster. This is done by calculating the cosine similarity between each token vector $\vec{u_i}$ and the centroids of the larger clusters and assigning the tokens to the closest large cluster. This re-clustering adapts the final number of clusters to the observed occurrences in the training data. Indeed, when there are few occurrences of a sense for a given ambiguous word type in the data, the SMT is likely not able to translate them properly due to the lack of training samples.

[5] code.google.com/archive/p/word2vec/

Finally, after clustering the training data, we use the centroids to assign each *new token from the test data* to a cluster, i.e. an abstract sense label, by selecting the closest centroid to it in terms of cosine distance in the embedding space.

3.3 Integration with Machine Translation

Our adaptive WSD system assigns a sense number for each ambiguous word token in the source-side of a parallel corpus. To pass this information to an SMT system, we use a factored phrase-based translation model (Koehn and Hoang, 2007). The factored model offers a principled way to supplement words with additional information – such as, traditionally, part-of-speech tags – without requiring any intervention in the translation tables. The features are combined in a log-linear way with those of a standard phrase-based decoder, and the goal remains to find the most probable target sentence for a given source sentence. To each source noun or verb token, we add a sense label obtained from our adaptive WSD system. To all the other words, we add a NULL label.[6] The translation system will thus take the source-side sense labels into consideration during the training and the decoding processes.

4 Datasets, Preparation and Settings

We evaluate our sense-aware SMT on the UN Corpus[7] (Rafalovitch and Dale, 2009) and on the Europarl Corpus[8] (Koehn, 2005). We select 0.5 million parallel sentences for each language pair from Europarl, as shown in Table 1. We also use the smaller WIT3 Corpus[9] (Cettolo et al., 2012), a collection of transcripts of TED talks, to evaluate the impact of costly model choices, namely the type of the resource (definition vs. examples), the length of the context window, and the k-means method (adaptive vs. original).

Before assigning sense labels, we first tokenize all the texts and identify the parts of speech (POS) using the Stanford POS tagger[10]. Then, we filter out the stopwords and the nouns which are proper names according to the Stanford Name Entity Recognizer[10]. Furthermore, we convert the plural forms of nouns to their singular form and the verb forms to infinitive using the stemmer and lemmatizer from NLTK[11], which is essential because WordNet has description entries only for singular nouns and infinitive form of verbs. The pre-processed text is used for assigning sense labels to each occurrence of a noun or verb which has more than one sense in WordNet.

For translation, we train and tune baseline and factored phrase-based models with Moses[12] (Koehn et al., 2007). We also carried out pilot experiments with neural machine translation (NMT). However, due to the large datasets NMT requires for training, its performance was below SMT on the datasets above, and sense labels did not improve it. We thus focus on SMT in what follows, and leave WSD for NMT for future studies.

We select the optimal model configuration based on the MT performance, measured with the traditional BLEU score (Papineni et al., 2002), on the WIT3 corpus for EN/ZH and EN/DE. Unless otherwise stated, we use the following settings in the k-means algorithm, starting from the implementation provided in Scikit-learn (Pedregosa et al., 2011):

- we use the definition of each sense for initializing the centroids in the adaptive k-means methods (and compare this later with using the examples);
- we set k_t equal to m_t, i.e. the number of senses of an ambiguous word type W_t;
- the window size for the context surrounding each occurrence is set to $c = 8$.

For the evaluation of intrinsic WSD performance, we use the V-metric, the F_1-metric, and their average, as used for instance at SemEval 2010 (Manandhar et al., 2010). To measure the impact of WSD on MT, besides BLEU, we also measure the actual impact on the nouns and verbs that appear in WordNet with several senses, by comparing how many of them are translated as in the reference translation, by our system vs. the baseline. For a certain set of tokens in the source data, we note as N_{improved} the number of tokens which are translated by our system as in the reference translation, but whose baseline translation differs from it. Conversely, we note as N_{degraded} the number of tokens which are translated by the

[6] In practice, these labels are simply appended to the tokens in the data following a vertical bar, e.g. 'rock|1' or 'great|NULL'.

[7] http://www.uncorpora.org/

[8] http://www.statmt.org/europarl/

[9] http://wit3.fbk.eu/

[10] http://nlp.stanford.edu/software/

[11] http://www.nltk.org/

[12] http://www.statmt.org/moses/

		Training		Development		Testing		Definitions		Examples	
		# lines	# tokens	# lines	# tokens	# lines	# tokens	# nouns	# verbs	# nouns	# verbs
EN/ZH	WIT3	150,000	3M	10,000	0.3M	50,000	1M	6,052	2,435	2,049	1,932
	UN	500,000	13M	5,000	0.14M	50,000	1.5M	8,165	3,382	2,810	2,716
EN/DE	WIT3	140,000	2.8M	5,000	0.16M	50,000	1M	8,308	2,384	3,662	2,042
	Europarl	500,000	14M	5,000	0.14M	50,000	1.4M	6,373	3,323	2,608	2,668
EN/FR	Europarl	∼	∼	∼	∼	∼	∼	8,279	4,022	2,276	2,054
EN/ES	Europarl	∼	∼	∼	∼	∼	∼	8,716	4,048	2,478	2,359
EN/NL	Europarl	∼	∼	∼	∼	∼	∼	8,667	4,023	2,439	2,318

Table 1: Statistics of the corpora used for machine translation: '∼' indicates a similar size, though not identical texts, because the English source texts for the different language pairs from Europarl are different. Hence, the number of words found in WordNet differ as well.

baseline system as in the reference, but differently by our system. We will use the normalized coefficient $\rho = (N_{\text{improved}} - N_{\text{degraded}})/T$, where T is the total number of tokens, as a metric focusing explicitly on the words submitted to WSD.[13]

5 Results

Using the data, settings, and metrics above, we investigate first the impact of two model choices on the performance: centroid initialization for k-means (definition or examples vs. random), and the length of the context window for each word. Then, we evaluate our adaptive clustering method on the WSD task, to estimate its intrinsic quality, and finally measure WSD+MT performance.

5.1 Optimal Values of the Parameters

5.1.1 Initialization of Adaptive k-means

We examine first the impact of the initialization of the sense clusters, on the WIT3 Corpus. In Table 2, we present the BLEU scores of our WSD+MT system in two conditions: when the k-means clusters are initialized with vectors from the definitions vs. from the examples provided in the WordNet synsets of ambiguous words. Moreover, we provide BLEU scores of baseline systems and oracle ones (i.e. using correct senses as factors), as well as the ρ score indicating the relative improvement of ambiguous words in our system wrt. the baseline. The use of definitions outperforms the use of examples, probably because there are more words with definitions than with examples in WordNet (twice as many, as shown in Table 1 in Section 4), but also because definitions may provide more helpful words to build the initial vectors, as they are more explicit than the examples.

[13]The values of N_{improved} and N_{degraded} are obtained using automatic word alignment. They do not capture, of course, the absolute correctness of a candidate translation, but only its identity or not with one reference translation.

All the values of ρ show clear improvements over the baseline, with up to 4% for DE/EN. As for the oracle scores, they outperform the baseline by a factor of 2–3 compared to our system.

Pair	Resource	BLEU			ρ (%)
		Baseline	Factored	Oracle	
EN/ZH	Definitions	15.23	**15.54**	16.24	+2.25
	Examples		15.41	15.85	+1.60
EN/DE	Definitions	19.72	**20.23**	20.99	+3.96
	Examples		19.98	20.45	+2.15

Table 2: Performance of our WSD+MT factored system for two language pairs from WIT3, with two initialization conditions for the k-means clusters, i.e. definitions or examples for each sense.

In addition, we compare the two initialization options above with random initializations of k-means clusters, in Table 3. To offer a fair comparison, we set the number of clusters, in the case of random initializations, respectively to the number of synsets with definitions or examples, for each word type. Clearly, our adaptive, informed initializations of clusters are beneficial to MT.

Resource	k-means initialization	
	Specific	Random
Definitions	**15.54**	15.34
Examples	**15.41**	15.27

Table 3: Performance of our WSD+MT factored system for EN-ZH from WIT3, comparing the two initialization conditions for the k-means clusters, i.e. definitions or examples for each sense, with random initializations.

5.1.2 Length of the Context Window

We investigate the effect of the size of the context window surrounding each ambiguous token, i.e. the number of words surrounding it that are considered for building its vector representation. Figure 3 displays the BLEU score of our WSD+MT

	System	V-score			F_1-score			Average			
		All	Nouns	Verbs	All	Nouns	Verbs	All	Nouns	Verbs	#clusters
Base.	MFS	0	0	0	64.85	57.00	72.70	32.42	29.50	25.40	1.00
	Random	4.40	4.60	4.20	32.35	30.60	34.10	18.45	17.60	19.30	4.00
	1ClusterPerIns	31.70	35.80	25.60	0.12	0.11	0.12	15.40	17.90	12.90	89.15
Top systems	Hermit (Jurgens and Stevens, 2010)	16.20	16.70	15.60	25.55	26.70	24.40	20.85	21.70	20.00	10.78
	UoY (Korkontzelos and Manandhar, 2010)	15.70	20.60	8.50	49.80	38.20	66.60	32.75	29.40	**37.50**	11.54
	KSU KDD (Elshamy et al., 2010)	15.70	18.00	12.40	36.90	24.60	54.70	26.30	21.30	33.50	17.50
	Duluth-WSI (Pedersen, 2010)	9.00	11.40	5.70	41.10	37.10	46.70	25.05	24.20	26.20	4.15
	Duluth-WSI-SVD-Gap (Pedersen, 2010)	0.00	0.00	0.10	63.30	57.00	72.40	31.65	28.50	36.20	1.02
	KCDC-PT (Kern et al., 2010)	1.90	1.00	3.10	61.80	56.40	69.70	31.85	28.70	36.40	1.50
	KCDC-GD (Kern et al., 2010)	6.90	5.90	8.50	59.20	51.60	70.00	33.05	28.70	39.20	2.78
	Duluth-Mix-Gap (Pedersen, 2010)	3.00	2.90	3.00	59.10	54.50	65.80	31.05	29.70	34.40	1.61
Ours	**Adaptive k-means + definition**	13.65	14.70	12.60	56.70	53.70	59.60	**35.20**	**34.20**	36.10	4.45
	Adaptive k-means + example	11.35	11.00	11.70	53.25	47.70	58.80	32.28	29.30	35.25	3.58

Table 4: WSD results from the SemEval 2010 shared task in terms of V-score, F_1 score and their average. Our adaptive k-means using definitions (last but one line) outperforms all the other systems on the average of V and F_1, when considering both nouns and verbs, or nouns only.

factored system when varying this size, on EN/ZH translation in the WIT3 Corpus, along with the (constant) score of the baseline. The performance of our system improves with the size of the window, reaching a peak around 8–10. This result highlights the importance of a longer context compared to the typical settings of SMT systems, which generally do not go beyond 6. It also suggests that MT systems which exploit effectively longer context, as we show here with a sense-aware factored MT system for ambiguous nouns and verbs, can significantly improve their lexical choice and their overall translation quality.

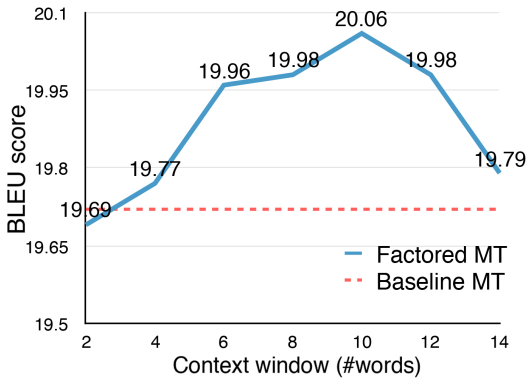

Figure 3: BLEU scores of our WSD+MT factored system on EN/ZH WIT3 data, along with the baseline score (constant), when the size of the context window around each ambiguous token (for building its context vector) varies from 2 to 14.

5.2 Word Sense Disambiguation Results

We evaluate in this section our WSD system on the dataset from the SemEval 2010 shared task (Manandhar et al., 2010), to assess how competitive it is, while acknowledging that our system uses external knowledge not available to SemEval participants.

Table 4 shows the WSD results in terms of V-score and F_1-score, comparing our method (bottom two lines) with other WSD systems that participated in SemEval 2010 (top four systems for each metric). We add three baselines provided by the task organizers for comparison: (1) Most Frequent Sense (MFS), which groups all occurrences of a word into one cluster, (2) 1ClusterPerInstance, which produces one cluster for each occurrence of a word, and (3) Random, which randomly assigns an occurrence to 1 out of 4 clusters (4 is the average number of senses from the ground-truth).

The V-score is biased towards systems generating a higher number of clusters than the number of gold standard senses. F_1-score measures the classification performance, i.e. how well a method assigns two occurrences of a word belonging to the same gold standard class. Hence, this metric favors systems that generate fewer clusters (for instance, if all instances were grouped into 1 cluster, the F_1-score would be high). As these two metrics are biased towards either small or large numbers of clusters, their average is a useful metric as well.

Table 4 shows that k-means initialized with definitions achieves high performance and ranks among the top systems for each metric individually, outperforming all other systems on the averaged metric (especially over nouns or all words). Moreover, the adaptive k-means method finds an

Language pair	Corpus	BLEU			ρ (%)
		Baseline	Factored	Oracle	
EN/ZH	UN	23.25	**23.69**	24.44	+2.26
EN/DE	Europarl	20.78	**21.32**	21.95	+1.57
EN/FR	Europarl	31.96	**32.20**	32.98	+1.21
EN/ES	Europarl	39.95	**40.37**	41.06	+1.04
EN/NL	Europarl	23.56	**23.84**	24.79	+1.38

Table 5: BLEU scores of our WSD+MT factored system, with both noun and verb senses, along with baseline MT and oracle WSD+MT, on five language pairs.

Language pair	Baseline	Factored (Nouns)				Factored (Verbs)			
		nouns		nouns + verbs	Oracle	verbs		nouns + verbs	Oracle
		BLEU	ρ (%)	ρ (%)		BLEU	ρ (%)	ρ (%)	
EN/ZH	23.25	**23.61**	+1.78	+1.93	24.05	**23.35**	+3.30	+3.14	24.17
EN/DE	20.78	**21.31**	+1.65	+1.48	21.45	**21.30**	+1.81	+1.79	21.87
EN/FR	31.96	**32.08**	+0.90	+0.82	32.36	**32.15**	+2.03	+2.13	32.98
EN/ES	39.95	**40.28**	+1.05	+0.96	40.59	**40.24**	+2.08	+1.15	41.06
EN/NL	23.56	**23.79**	+1.13	+0.87	24.05	**23.70**	+2.58	+2.71	24.46

Table 6: BLEU scores of our WSD+MT factored system, trained separately on disambiguated nouns vs. verbs, and tested separately or jointly, along with baseline MT and oracle WSD+MT, on five language pairs.

average number of senses of 4, which is close to the ground-truth value provided by SemEval (4.46). These results show that our method, despite its simplicity, is effective and provides competitive performance against prior art, partly thanks to additional knowledge not available to the shared task systems.

5.3 Machine Translation Results

Table 5 displays the performance of our factored MT systems trained with noun and verb senses on five language pairs by using the dataset mentioned in Table 1. Our system performs consistently better than the MT baseline on all pairs, with the largest improvements achieved on EN/ZH and EN/DE. To better understand the improvements over the baseline MT, we also provide the BLEU score of an oracle system which has access to the reference translation of the ambiguous words through the alignment provided by GIZA++. According to the results, our factored MT system bridges around 40% of the gap between the baseline MT system and the oracle system on EN/DE and 30% on EN/ZH.

As shown in Table 6, the translation quality of our factored MT outperforms the baseline when trained with either noun senses or verb senses separately. However, in some cases, our factored MT system trained with both noun and verb senses performs worse than with noun and verb senses separately. This may be due to the lack of sufficient training data to learn reliably using all the additional factors – as we observed when training on the smaller WIT3 Corpus.

Lastly, Table 7 shows the confusion matrix for our factored MT and the baseline MT systems when comparing the reference translation of nouns and verbs separately, using GIZA++ alignment. In particular, the confusion matrix displays the number of labeled tokens which are translated as in the reference or not ('Correct' vs. 'Incorrect'). As we can observe, the number of tokens that our factored MT system translates correctly while the baseline MT does not, is two times largers than the number of tokens that the baseline MT system finds correctly while our factored MT does not.

6 Conclusion

We presented a sense-aware statistical MT system which uses a larger context than standard ones, through an adaptive context-dependent k-means clustering algorithm for WSD. The algorithm utilizes semantic information from WordNet to identify the dominant clusters, which correspond to senses in the source side of a parallel corpus. The proposed adaptive k-means method is straightforward, yet it provides competitive WSD performance on data from the SemEval 2010 shared task. For MT, our experiments with five language pairs show that our sense-aware MT system consistently improves over the baseline. As future work, we plan to integrate sense information for ambiguous words to neural MT and investigate

		Factored (Nouns)				Factored (Verbs)			
		nouns		nouns + verbs		verbs		nouns + verbs	
		Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
EN/ZH	Correct	138,876	4,402	138,264	5,075	37,132	1,166	36,647	1,527
Baseline	Incorrect	8,454	75,690	9,472	74,541	3,939	41,728	4,149	41,077
EN/DE	Correct	91,966	1,473	91,376	2,035	18,370	664	18,214	812
Baseline	Incorrect	4,268	71,037	4,525	69,931	1,892	47,105	2,029	46,795

Table 7: Detailed confusion matrix of our factored MT system and the baseline MT system with respect to the reference on the EN/DE pair from Europarl corpus and the EN/ZH from UN corpus.

other effective ways to enable access to longer context.

Acknowledgments

We are grateful for their support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (grant n. 147653, see www.idiap.ch/project/modern/) and to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see www.summa-project.eu). We thank the reviewers for their helpful suggestions.

References

Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Science & Business Media.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece, pages 33–41.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, Ohio, pages 763–770.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pages 9–16.

Clara Cabezas and Philip Resnik. 2005. Using WSD techniques for lexical selection in statistical machine translation. Technical report, DTIC Document.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of Association for Computational Linguistics*. Michigan, USA, pages 387–394.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 61–72.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pages 33–40.

Wesam Elshamy, Doina Caragea, and William H Hsu. 2010. KSU KDD: Word sense induction by clustering in topic space. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010)*. Association for Computational Linguistics, Los Angeles, California, pages 367–370.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, USA.

David Jurgens and Keith Stevens. 2010. Hermit: Flexible clustering for the Semeval-2 WSI task. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010))*. Association for Computational Linguistics, Los Angeles, California, pages 359–362.

Roman Kern, Markus Muhr, and Michael Granitzer. 2010. KCDC: Word sense induction by using grammatical dependencies and sentence phrase structure. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010)*. Association for Computational Linguistics, Los Angeles, California, pages 351–354.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit*. Phuket, Thailand, pages 79–86.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 868–876.

Philipp Koehn, Hieu Hoang, Birch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting*

of the Association for Computational Linguistics. pages 177–180.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Uoy: Graphs of ambiguous vertices for word sense induction and disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010)*. Los Angeles, California, pages 355–358.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA, pages 281–297.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan. 2010. SemEval-2010 task 14: Word sense induction and disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010)*. Association for Computational Linguistics, Los Angeles, California, pages 63–68.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Steven Neale, Luıs Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. 2016. Word sense-aware machine translation: Including senses as contextual features for improved translation models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portoroz, Slovenia, pages 2777–2783.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*. Philadelphia, USA, pages 311–318.

Ted Pedersen. 2010. Duluth-WSI: SenseClusters applied to the sense induction task of SemEval-2. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval-2010)*. Association for Computational Linguistics, Los Angeles, California, pages 363–366.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Alexandre Rafalovitch and Robert Dale. 2009. United Nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*. Ontario, Canada, pages 292–299.

Jinsong Su, Deyi Xiong, Shujian Huang, Xianpei Han, and Junfeng Yao. 2015. Graph-Based collective lexical selection for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 1238–1247.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada, pages 771–778.

Deyi Xiong and Min Zhang. 2014. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore MD, USA, pages 1459–1469.