Word Sense Induction with Neural biLM and Symmetric Patterns

Asaf Amrami † and **Yoav Goldberg** † ‡

† Computer Science Department, Bar Ilan University, Israel
‡ Allen Institute for Artificial Intelligence

{asaf.amrami, yoav.goldberg}@gmail.com

Abstract

An established method for Word Sense Induction (WSI) uses ~~a language model~~ to predict probable substitutes for target words, and induces senses by clustering these resulting substitute vectors.

We replace the ngram-based language model (LM) with a recurrent one. Beyond being more accurate, the use of the recurrent LM allows us to effectively query it in a creative way, using what we call *dynamic symmetric patterns*. The combination of the RNN-LM and the dynamic symmetric patterns results in strong substitute vectors for WSI, allowing to surpass the current state-of-the-art on the SemEval 2013 WSI shared task by a large margin.

1 Introduction

We deal with the problem of *word sense induction* (WSI): given a target lemma and a collection of within-sentence usages it, cluster the usages (**instances**) according to the different senses of the target lemma. For example, for the sentences:

(a) We spotted a large ***bass*** in the ocean.

(b) The ***bass*** player did not receive the acknowledgment she deserves.

(c) The black sea ***bass***, is a member of the wreckfish family.

We would like to cluster (a) and (c) in one group and (b) in another.[1] Note that some mentions are ambiguous. For example, (d) matches both the music and the fish senses:

(d) ***Bass*** scales are the worst.

This calls for a *soft clustering*, allowing to probabilistically associate a given mention to two senses.

The problem of WSI has been extensively studied with a series of shared tasks on the topic (Agirre and Soroa, 2007; Manandhar et al., 2010; Jurgens and Klapaftis, 2013), the latest being SemEval 2013 Task 13 (Jurgens and Klapaftis, 2013). Recent state-of-the-art approaches to WSI rely on generative graphical models (Lau et al., 2013; Wang et al., 2015; Komninos and Manandhar, 2016). In these works, the sense is modeled as a latent variable that influences the context of the target word. The later models explicitly differentiate between local (syntactic, close to the disambiguated word) and global (thematic, semantic) context features.

Substitute Vectors Baskaya et al. (2013) take a different approach to the problem, based on *substitute vectors*. They represent each instance as a distribution of possible substitute words, as determined by a language model (LM). The substitute vectors are then clustered to obtain senses.

Baskaya et al. (2013) derive their probabilities from a 4-gram language model. Their system (AI-KU) was one of the best performing at the time of SemEval 2013 shared task. Our method is inspired by the AI-KU use of substitution based sense induction, but deviate from it by moving to a recurrent language model. Besides being more accurate, this allows us to further improve the quality of the derived substitutions by the incorporation of *dynamic symmetric patterns.*

BiLM Bidirectional RNNs were shown to be effective for word-sense disambiguation and lexical substitution tasks (Melamud et al., 2016; Yuan et al., 2016; Raganato et al., 2017). We adopt the ELMo biLM model of Peters et al. (2018), which

[1] This example shows *homonymy*, a case where the same word form has two distinct meaning. A more subtle case is *polysemy*, where the senses share some semantic similarity. In "She played a low bass note", the sense of *bass* is related to the sense in (b), but distinct from it. The WSI task we tackle in this work deals with both cases.

was shown to produce very competitive results for many NLP tasks. We use the pre-trained ELMo biLM provided by Peters et al. (2018).[2] However, rather than using the LSTM state vectors as suggested in the ELMo paper, we opt instead to use the predicted word probabilities. Moving from continuous and opaque state vectors to discrete and transparent word distributions allows far better control of the resulting representations (e.g. by sampling, re-weighting and lemmatizing the words) as well as better debugging opportunities.

As expected, the move to the neural biLM already outperforms the AI-KU system, and matches the previous state-of-the-art. However, we observe that the substitute vectors do not take into account the disambiguated word itself. We find that this often results in noisy substitutions. As a motivating example, consider the sentence "the doctor recommends *oranges* for your health". Here, *running* is a perfectly good substitution, as the "fruitness" of the target word itself isn't represented in the context. We would like the substitutes word distribution representing the target word to take both kinds of information—the context as well as the target word—into account.

Dynamic Symmetric Patterns Our main proposal incorporates such information. It is motivated by Hearst patterns (Hearst, 1992; Widdows and Dorow, 2002; Schwartz et al., 2015), and made possible by neural LMs. Neural LMs are better in capturing long-range dependencies, and can handle and predict unseen text by generalizing from similar contexts. Conjunctions, and in particular the word *and*, are known to combine expressions of the same kind. Recently, Schwartz et al. (2015) used conjunctive symmetric patterns to derive word embeddings that excel at capturing word similarity. Similarly, Kozareva et al. (2008) search for doubly-anchored patterns including the word *and* in a large web-corpus to improve semantic-class induction. The method of Schwartz et al. (2015) result in context-independent embeddings, while that of Kozareva et al. (2008) takes some context into account but is restricted to exact corpus matches and thus suffers a lot from sparsity.

We make use of the rich sequence representation capabilities of the neural biLM to derive *context-dependent symmetric pattern substitutions*. Relying on the generalization properties of neural language models and the abundance of the "X and Y" pattern, we present the language model with a dynamically created incomplete pattern, and ask it to predict probable completion candidates. Rather than predicting the word distribution following *the doctor recommends* __, we instead predict the distribution following *the doctor recommends oranges and* __. This provides substantial improvement, resulting in state-of-the-art performance on the SemEval 2013 shared task.

The code for reproducing the experiments and our analyses is available at `https://github.com/asafamr/SymPatternWSI`.

2 Method

Given a target word (lemma and its part-of-speech pair), together with several sentences in which the target word is used (instances), our goal is to cluster the word usages such that each cluster corresponds to a different sense of the target word. Following the SemEval 2013 shared task and motivating example (d) from the introduction, we seek a soft (probabilistic) clustering, in which each word instance is assigned with a probability of belonging to each of the sense-clusters.

Our algorithm works in three stages: (1) We first associate each instance with a probability distribution over in-context word-substitutes. This probability distribution is based on a neural biLM (section 2.1). (2) We associate each instance with k representatives, each containing multiple samples from its associated word distributions (section 2.3). (3) Finally, we cluster the representatives and use the hard clustering to derive a soft-clustering over the instances (section 2.4).

We use the pre-trained neural biLM as a black-box, but use linguistically motivated processing of both its input and its output: we rely on the generalization power of the biLM and query it using *dynamic symmetric patterns* (section 2.2); and we lemmatize the resulting word distributions.

Running example In what follows, we demonstrate the algorithm using a running example of inducing senses from the word *sound*, focusing on the instance sentence:

*I liked the **sound** of the harpsichord.*

2.1 biLM Derived Substitutions

We follow the ELMo biLM approach (Peters et al., 2018) and consider two separately

[2] We thank the ELMo team for sharing the pre-trained models.

trained language models, a forward model trained for predicting $p_{\rightarrow}(w_i|w_1, ..., w_{i-1})$ and a backward model $p_{\leftarrow}(w_i|w_n, ..., w_{i+1})$. Rather than combining the two models' predictions into a single distribution, we simply associate the target word with two distributions, one from $p_{\rightarrow}$ and one from $p_{\leftarrow}$. For convenience, we use $LM_{\rightarrow}(w_1 w_2 ... w_{i-1}\_\_)$ to denote the distribution $p_{\rightarrow}(w_i|w_1, ..., w_{i-1})$ and $LM_{\leftarrow}(\_\_ w_{i+1} w_{i+2} ... w_n)$ to denote $p_{\leftarrow}(w_i|w_n, ..., w_{i+1})$.

Context-based substitution In the purely context-based setup (the one used in the AI-KU system) we represent the target word *sounds* by the two distributions:

$LM_{\rightarrow}$(<s> I liked the __)
$LM_{\leftarrow}$(__ of the harpsichord </s>)

The resulting top predictions from each distribution are:
{*idea:0.12, fact:0.07, article: 0.05, guy: 0.04, concept: 0.02*} and
{*sounds:0.04, version: 0.03, rhythm: 0.03, strings: 0.03, piece: 0.02*} respectively.

2.2 Dynamic Symmetric Patterns

As discussed in the introduction, conditioning solely on context is ignoring valuable information. This is evident in the resulting word distributions. We use the coordinative symmetric pattern *X and Y* in order to produce a substitutes vector incorporating both the word and its context. Concretely, we represent a target word w_i by $p_{\rightarrow}(w'|w_1, ..., w_i, \texttt{and})$ and $p_{\leftarrow}(w'|w_n, ..., w_i, \texttt{and})$. For our running example, this translates to:

$LM_{\rightarrow}$(<s> I liked the sound and __)
$LM_{\leftarrow}$(__ and sound of the harpsichord . </s>)

with resulting top words: {*feel: 0.15, felt: 0.11, thought: 0.07, smell: 0.06, sounds: 0.05*} and
{*sight: 0.16, sounds: 0.11, rhythm: 0.04, tone: 0.03, noise: 0.03*}.

The distributions predicted using the *and* pattern exhibit a much nicer behavior, and incorporate global context (resulting in sensing related substitutes) as well as local and syntactic information that resulting from the target word itself. Table 1 compares the context-only and symmetric-pattern substitutes for two senses of the word *sound*.

2.3 Representative Generation

To perform fuzzy clustering, we follow AI-KU and associate each instance with k representatives, but deviate in the way the representatives are generated. Specifically, each representative is a set of size 2ℓ , containing ℓ samples from the forward distribution and ℓ samples from the backward distribution. In the symmetric pattern case above, a plausible representative, assuming $\ell = 2$, would be: {*feel, sounds, sight, rhythm*} where two words were predicted by each side LM. In this work, we use $\ell = 4$ and $k = 20$.

2.4 Sense Clustering

After obtaining k representatives for each of the n word instances, we cluster the nk representatives into distinct senses and translate this hard-clustering of representatives into a probabilistic clustering of the originating instances.

Hard-clustering of representatives Let V be the vocabulary obtained from all the representatives. We associate each representative with a sparse $|V|$ dimensional bag-of-features vector, and arrange the representatives into a $nk \times |V|$ matrix M where each row corresponds to a representative. We now cluster M's rows into senses. We found it is beneficial to transform the matrix using TF-IDF. Treating each representative as a document, TF-IDF reduces the weight of uninformative words shared by many representatives. We use agglomerative clustering (cosine distance, average linkage) and induce a fixed number of clusters.[3] We use `sklearn` (Pedregosa et al., 2011) for both TF-IDF weighting and clustering.

Inducing soft clustering over instances After clustering the representatives, we induce a soft-clustering over the instances by associating each instance j to sense i based on the proportion of representatives of j that are assigned to cluster i.

2.5 Additional Processing

Lemmatization The WSI task is defined over lemmas, and some target words have morphological variability within a sense. This is especially common with verb tenses, e.g., "I **booked** a flight" and "I am **booking** a flight". As the conjunctive

[3] In this work, we use 7 clusters, which roughly matches the number of senses for each target word in the corpus. Dynamically selecting the number of clusters is left for future work. The effect of changing the number of clusters is explored in the supplementary material.

Context Only		Symmetric Pattern	
Forward dist.	Backward dist.	Forward dist.	Backward dist.
This is a ***sound*** idea, I like it.			
sad 0.02	bad 0.12	welcome 0.09	funny 0.10
great 0.02	good 0.09	practical 0.03	beautiful 0.05
huge 0.02	great 0.06	comprehensive 0.03	fun 0.04
very 0.02	wonderful 0.05	light 0.02	simple 0.04
lesson 0.02	nice 0.04	balanced 0.02	interesting 0.03
I liked the ***sound*** of the harpsichord			
idea 0.12	sounds 0.04	feel 0.15	sight 0.16
fact 0.07	version 0.03	felt 0.11	sounds 0.11
article 0.05	rhythm 0.03	thought 0.07	rhythm 0.04
guy 0.04	strings 0.03	smell 0.06	tone 0.03
concept 0.02	piece 0.03	sounds 0.05	noise 0.03

Table 1: Predicted substitutes for two senses of sound, for context-only and the symmetric-pattern approaches.

symmetric pattern favors morphologically-similar words, the resulting substitute vectors for these two sentences will differ, each of them agreeing with the tense of its source instance. To deal with this, we lemmatize the predictions made by the language model prior to adding them to the representatives. Such removal of morphological inflection is straightforward when using the word distributions but much less trivial when using raw LM state vectors, further motivating our choice of working with the word distributions. The substantial importance of the lemmatization is explored in the ablation experiments in the next section, as well as in the supplementary material.

Distribution cutoff and bias Low ranked LM prediction tend to become noisier. We thus consider only the top 50 word predicted by each LM, re-normalizing their probabilities to sum to one. Additionally, we ignore the final bias vector during prediction (words are predicted via $softmax(Wx)$ rather than $softmax(Wx + b)$). This removes unconditionally probable (frequent) words from the top LM predictions.

3 Experiments and Results

We evaluate our method on the SemEval 2013 Task 13 dataset (Jurgens and Klapaftis, 2013), containing 50 ambiguous words each with roughly 100 in-sentence instances, where each instance is soft-labeled with one or more WordNet senses.

Experiment Protocol Due to the stochastic nature of the algorithm, we repeat each experiment 30 times and report the mean scores together with the standard deviation.

Evaluation metrics We follow previous work (Wang et al., 2015; Komninos and Manandhar, 2016) and evaluate on two measures: ***Fuzzy Normalized Mutual Information (FNMI)*** and ***Fuzzy B-Cubed (FBC)*** as well as their geometric mean (**AVG**).

Systems We compare against three graphical-model based systems which, as far as we know, represent the current state of the art: **MCC-S** (Komninos and Manandhar, 2016), **Sense-Topic** (Wang et al., 2015) and **unimelb** (Lau et al., 2013). We also compare against the **AI-KU** system. Wang et al. also present a method for dataset enrichment that boosted their model performance. We didn't use the suggested methods and compare ourselves to the vanilla settings, but report the enrichment numbers as well.

Results Table 2 summarizes the results. Our system using symmetric patterns outperforms all other setups with an AVG score of 25.4, establishing a new state-of-the-art on the task.

Ablation and analysis We perform ablations to explore the contribution of the different components (Symmetric Patterns (SP), Lemmatization (LEM) and TF-IDF re-weighting). Figure (1) shows the results for the entire dataset (ALL, top), as well as broken-down by part-of-speech. All components are beneficial and are needed for obtaining the best performance in all cases. However, their relative importance differs across parts-of-speech. Adjectives gain the most from the use of the dynamic symmetric patterns, while nouns

Model	FNMI	FBC	AVG
Original task dataset			
Ours	**11.26** $\pm$ 0.48	57.49 $\pm$ 0.23	**25.43** $\pm$ 0.48
MCC-S	7.62	55.6	20.58
Sense-Topic (SW)	7.14	55.4	19.89
Sense-Topic	6.96	53.5	19.30
AI-KU	6.5	39.0	15.92
unimelb	6.0	48.3	17.02
With data enrichment			
Sense-Topic (AAC)	9.39	**59.1**	23.56
Sense-Topic (AUC)	9.74	54.5	23.04

Table 2: Evaluation Results on the SemEval 2013 Task 13 Dataset. SW: Embeddings similarity based feature weighting. AAC: Extending instance sentences from their traced source. AUC: Adding similar sentences from the dataset originating corpus. We report our mean scores over 30 runs $\pm$ standard deviation

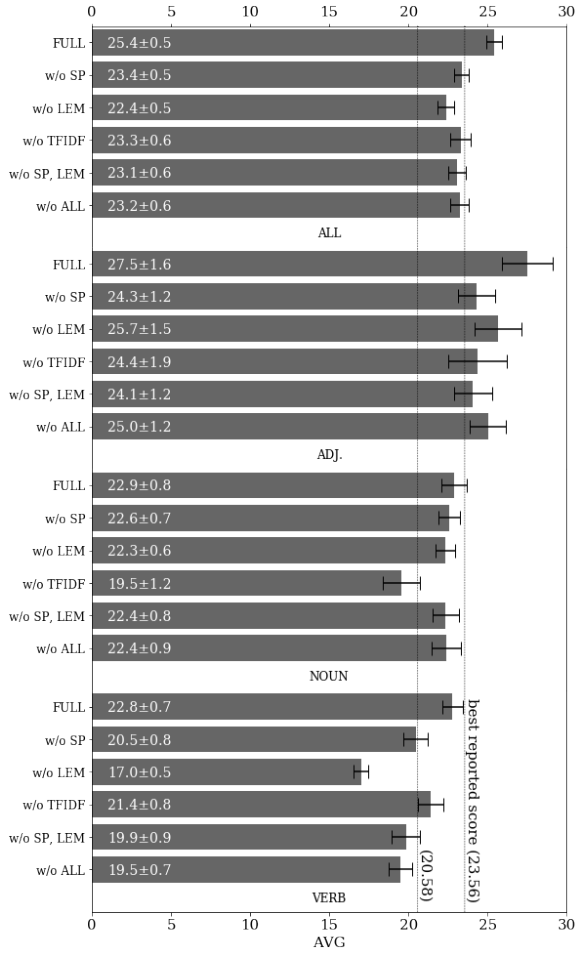

Figure 1: Ablation break down by part of speech, each part of speech was averaged across run. Bars are mean of means and error bars are standard deviations.

gain the least. For verbs, the lemmatization is crucial for obtaining good performance, especially when symmetric patterns are used: using symmetric patterns without lemmatization, the mean score drops to 17.0. Lemmatization without symmetric patterns achieves a higher mean score of 20.5, while using both yields 22.8. Finally, for nouns it is the TF-IDF scoring that plays the biggest role.

4 Conclusions

We describe a simple and effective WSI method based on a neural biLM and a novel dynamic application of the *X and Y* symmetric pattern. The method substantially improves on the state-of-the-art. Our results provide further validation that RNN-based language models contain valuable semantic information.

The main novelty in our proposal is querying the neural LM in a creative way, in what we call *dynamic symmetric patterns*. We believe that the use of such dynamic symmetric patterns (or more generally *dynamic Hearst patterns*) will be beneficial to NLP tasks beyond WSI.

In contrast to previous work, we used discrete predicted word distributions rather than the continuous RNN states. This paid off by allowing us to inspect and debug the representation, as well to control it in a meaningful way by injecting linguistic knowledge in the form of lemmatization, and by distributional cutoff and TF-IDF re-weighing. We encourage others to consider using explicit, discrete representations when appropriate.

Acknowledgments The work was supported in part by the Israeli Science Foundation (grant number 1555/15 and the German Research Foundation via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1).

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 7–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Osman Baskaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 300–306.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 290–299.

Alexandros Komninos and Suresh Manandhar. 2016. Structured generative models of continuous features for word sense induction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3577–3587.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, pages 1048–1056, Columbus, Ohio. Association for Computational Linguistics.

Jey Han Lau, Paul Cook, and Timothy Baldwin. 2013. unimelb: Topic modelling-based word sense induction. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 307–311.

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

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 258–267.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D Ziebart, and T Yu Clement. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association of Computational Linguistics*, 3(1):59–71.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Supplementary Material

Statistics of the SemEval 2013 Task 13 Dataset

SemEval 2013 Task 13 consists of 50 targets, each has a lemma and a part of speech (20 verbs, 20 nouns and 10 adjectives). We use the dataset only for evaluation. Most targets have around 100 labeled instances (sentences containing a usage of the target in its designated part of speech together with one or more WordNet senses assigned by human labeler). Exceptions are the targets of trace.n and book.v which have 37 and 22 labeled instances accordingly. Leaving out the two anomalous targets mentioned above we are left with 4605 instances from 48 targets: 19 verb, 19 noun and 10 adjective targets. We note that the small size of the dataset should make one cautious to draw quick conclusions, yet, our results seem to be consistent.

Effect of the Choice of Number of Clusters

An important statistic of the dataset is the number of senses per target. The average number of senses per target in the dataset is 6.94 (stdev:2.71). Breaking down by part of speech, the means and standard deviations of target senses are: verbs: 5.90 ($\pm$1.37), nouns: 7.32 ($\pm$2.21), adjectives: 7.11 ($\pm$3.54). In this work we follow this statistic and always look for 7 clusters. Figure 2 shows the accuracy as a function of the number of clusters. While 7 clusters indeed produces the highest scores, all numbers in the range 4 to 15 produce state-of-the-art results. We leave the selection of per-instance number of clusters to future work.

Figure 2 also tells us our system is better at inducing senses for adjectives, at least according to task score.

The Importance of Lemmatization

The ablation results in the paper indicate that for verbs, using symmetric patterns without lemmatization yields poor results. We present the analysis the motivated our use of lemmatization. Consider the samples from the biLM with and without symmetric patterns, for the instance *It was when I was a high-school student that I* ***became*** *convinced of this fact for the first time.*

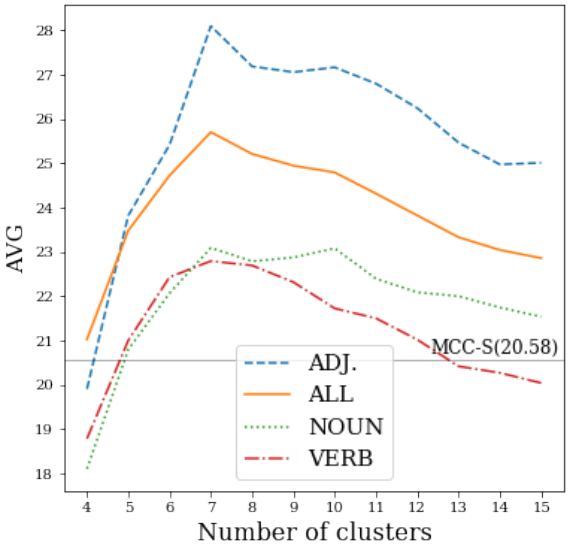

Figure 2: AVG score by number of clusters.

fw LM, no SP:	didn, write, 'd, learnt, start
bw LM, no SP:	seem, be, grow, be, be
fw LM, with SP:	went, got, started, wasn, loved
bw LM, with SP:	1990s, decade, 1980s, afterwards, changed

Another sentence, in another tense: *The issue will* ***become*** *more pressing as an estimated 40,000 to 50,000 Chinese, mostly unskilled, come to settle each year.*

fw LM, no SP:	be, be, remain, likely, be
bw LM, no SP:	becoming, grown becoming, much, becomes
fw LM, with SP:	remains, remain, which, continue, how
bw LM, with SP:	rising, overseas, booming, abroad, expanded

When using the symmetric patterns, the predicted verbs tend to share the tense of the target word.

This results in targets of different tenses having nearly distinct distributions, even when the targets share the same sense, splitting the single sense cluster to two (or more) tense clusters. We quantify this intuition by computing the correlation between tense and induced clusters (senses), as given by the Normalized Mutual Information (NMI). We measure NMI between verb instance tense in sentence and their most probable induced cluster in the different settings, as well as the NMI of the verb instances and the gold clusters. Table 3 summarize the results. We see that in the gold clusters there is indeed very little correlation

(0.15) between the the tense and the sense. When using SP but not lemmatization (w/o LEM), the correlation is substantially higher (0.67). When not using either lemmatization of SP (w/o LEM and SP) the correlation is 0.27, much closer to the gold one. Performing explicit lemmatization naturally reduces the correlation with tense, and using the full model (Final model) results in a correlation to 0.22, close to the gold number of 0.15.

Some Failure Modes of Dynamic Symmetric Patterns

While the use of dynamic symmetric patterns improves performance and generally produces good substitutes for contextualized words, we also identify some failure modes and unexpected behavior.

Common phrases involving conjunctions Some target words have a strong prior to appear in common phrases involving a conjunction, causing the strong local pattern to override context-based hints. For example, when the LM is asked to complete ... *state and* __, its prior on *church* makes it a very probable completion, regardless of context and sense. This phenomena motivated our use TF-IDF for weighing of too common words. Relatedly, a common completion for symmetric patterns is the word *then*, as *and then* is a very common phrase. This completion even ignores the target word and could be troublesome if a global, cross-lemma, clustering is attempted.

Multi word phrases substitutes Sometime the LM does interpret the *and* as a trigger for a symmetric relation, but on a chunk extending beyond the target word. For example, when presented with the query *The human heart not only makes heart sounds and* __, the forward LM predicted in its top twenty suggestions the word *muscle*, followed by a next-word prediction of *movements*. That is, the symmetry extends beyond "sounds" to the phrase "heart sounds" which could be substitutes by "muscle movements". We didn't specifically address this in the current work, but note that restricting the prediction to agree with the target word on part-of-speech and plurality may help in mitigating this. Furthermore, this suggests an exciting direction for moving from single words towards handling of multi-word units.

Settings	NMI (mean $\pm$ STD)
Gold labels	0.15 ± 0.07
Final model	0.22 ± 0.12
w/o SP	0.19 ± 0.08
w/o TFIDF	0.18 ± 0.07
w/o LEM	$\mathbf{0.67 \pm 0.12}$
w/o LEM and SP	0.26 ± 0.09
w/o ALL	0.24 ± 0.08

Table 3: Correlation between tense and sense. NMI is averaged on all verbs, using best matching sense. SP: Symmetric Patterns, LEM: Lemmatizing predictions, ALL: LEM, SP, TFIDF. The bold line show symmetric patterns without lemmatization excessively correlates tense and sense and provides additional validation to our hypothesis, suggesting its essential to lemmatizate when symmetric patterns are used.