LOW-DIMENSIONAL CONTEXT-DEPENDENT TRANSLATION MODELS

Avneesh Singh Saluja
Thesis Proposal, October 2014

Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Chris Dyer
Ian Lane
Ying Zhang
Hal Daumé III

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

1 Introduction & Motivation

Context matters when modeling translation. We have seen empirical evidence of this assertion in the transition from word- to phrase-based models, which achieve start-of-the-art performance by relying on long multiword units or phrases to incorporate local context (Koehn et al., 2003, Chiang, 2007). More recently, neural and factored translation models have started to condition on large amounts of source-language context and have reported sizable gains in translation performance (Feng and Cohn, 2013, Devlin et al., 2014). Like the current zeitgeist, this thesis takes seriously the premise that modeling contextual dependencies in machine translation (MT) is key to effective translation. However, state-of-the-art approaches predominantly model these dependencies via larger translation units. Larger units mean larger models, resulting in problems in computational efficiency (runtime and memory) and statistical efficiency (how can we reliably learn parameters for billions of rules from millions of sentences?). Furthermore, we contend that translation models should be sensitive to far more than "local" context: larger contexts are observed during evaluation (e.g., the entire sentence to be translated or even the document containing the sentence), and this information should be used when translating since it is computationally inexpensive to do so and produces better translations (§3). For example, a subject could be separated by a long subordinate clause and if we consider limited local context, the subject's influence on a subsequent verb phrase may be ignored. But if we cannot reliably estimate context-insensitive models or effectively decode with them, how should we hope to develop models that are sensitive to the rich input context available during test time? Clearly, another approach is necessary.

The thesis statement and central hypothesis of this work is that *while context influences translation, its influence is inherently low-dimensional, and problems of computational and statistical tractability can be solved by using dimensionality reduction and representation learning techniques.* To make the right translation choices for ambiguous words like *bank* or *watch*, it is unnecessary to embed the word in a long phrase as a means of incorporating context (by naive memorization of sequences); a small amount of information, e.g., the presence of a determiner before *watch*, or the knowledge that *bank* occurs amongst context that is semantically related to finance, suffices. The low-dimensional representations we recover intuitively capture this observation, that the phenomena that drive translation are controlled by context residing in a more compact space than the lexical-based (word or n-gram) "one-hot" or count-based spaces. Furthermore, they allow more reliable parameter estimates from less data in a more computationally efficient manner (Kakade and Foster, 2007).

Outside of the prevalent yet naive way of incorporating small amounts of local context by enlarging or heavily lexicalizing translation rules, there has been surprisingly limited amounts of work in context-based disambiguation for MT.[1] Chan et al. (2007) and Carpuat (2008) were the first to transfer ideas from word-sense disambiguation (WSD) into MT by formulating WSD in terms of ranking translation options in a target language given the input source sentence (later referred to as phrase sense disambiguation).[2] Features are extracted from the source sentence only, and this idea has also been used in approaches (Stroppa et al.,

[1] Although interestingly Berger et al. (1996) motivate the application of maximum entropy models in NLP with a translation sense disambiguation task.

[2] Some of these ideas have subsequently been used for specific applications in MT, e.g., suggesting translations for OOV words (Daumé III and Jagarlamudi, 2011) or identifying when words obtain new senses due to a change in domain (Carpuat et al., 2013).

2007, Gimpel and Smith, 2008, He et al., 2008) that enrich the standard linear translation setting by adding source context features and tuning their weights, along with the weights of standard MT features like the relative phrasal frequency estimates and the language model, using standard algorithms like MERT (Och, 2003). While some of these approaches have been effective (potentially due to the intrinsic dimensionality of context), this line of work has in general achieved mixed results because this low-dimensional view is never explicitly leveraged. Most of these models are heavily dependent on high-dimensional lexical features and manually-defined coarser features like part-of-speech (POS) tags, and minimal effort is made to reason about context in a lower-dimensional space or about translation rules (in terms of context) in such spaces.

A second line of attack, namely n-gram-based machine translation (Mariño et al., 2006, Durrani et al., 2011), uses Markov models to model context in the form of a history. The n-gram framework allows the use of heuristic smoothing techniques from language modeling (Chen and Goodman, 1999) to indirectly capture context low-dimensionally, and while more principled approaches based on Pitman-Yor priors achieve good performance (Feng and Cohn, 2013), the n-gram methods are still limited by their unidirectional (i.e, left-to-right) notion of context and their reliance on smoothing as a proxy to reasoning about the effect of context dimensionality on translation.

In most of these previous approaches, global context adaptation is carried out on top of a massive phrase-based model. Due to the sheer number of phrases to consider, this decision often limits the extent to which context can be considered and for tractability reasons, context disambiguation often boils down to a lexical selection problem whereas ideally we should consider translation rules. Despite empirical evidence showing models with composed rules (ones that can be formed out of smaller rules) outperform their minimal counterparts by weakening independence assumptions in the translation model (Koehn et al., 2003, Galley et al., 2006), we propose to work with minimal translation rules as our basic units. The reasoning is that if we have better translation models that take into account context more effectively than including limited amounts within translation rules, we can eliminate the gigantic grammars that phrase-based translation has relied on and work with smaller, simpler models that generalize better to new corpora. Thus, a secondary objective of this thesis is to explore the performance of minimal grammars in conjunction with low-dimensional context-dependent models, and compare them to composed grammars which incorporate local context dependence within rules. This motivation is in line with Vaswani et al. (2011), who leverage the n-gram translation framework to explore this transition from minimal to composed grammars in the setting of tree-to-string translation (Huang et al., 2006).

In this thesis, we consider low-dimensional representations of context, as well as low-dimensional representations of translation units that are expressed (featurized) in terms of context. Specifically, three instantiations of the low-dimensional assumption are explored:

1. **Low-dimensional embeddings of translation units** (§2, 90% complete): we propose a latent-variable model for synchronous context-free grammars (SCFGs) and apply it to hierarchical phrase-based translation (HPBT). The non-terminals in each rule are augmented with latent states in a context-dependent manner, which we learn from a parallel corpus. Specifically, non-terminals are refined by expressing them in terms of low-dimensional projections of the context in which they occur. In this case, we project a high-dimensional representation of a translation rule, represented with the empirical

covariances of inside and outside tree features in synchronous trees where the rule occurs as a non-terminal, into a low-rank space.

2. **Low-dimensional embeddings of context** (§3, 0% complete): we adopt the multi-view assumption (Foster et al., 2008, Dhillon et al., 2011), which states that we can use two complementary views of the data (e.g., the left and right contexts a source phrase occurs in) to recover a low-dimensional basis via canonical correlation analysis (CCA). Supervised learning can then proceed in this low-rank space but with reduced sample complexity. The intuition is to utilize information in both views which correlate well with each other. In this case, we project a high-dimensional representation of context defined in terms of lexical and syntactic features into a low-dimensional space using CCA. We propose to model phrasal choice and translation sense disambiguation by conditioning on low-dimensional representations of extreme source context in this manner, taking into account both local (sentence-specific) and global (paragraph or document-specific) information.

3. **Low-dimensional embeddings and semi-supervised learning** (§4, 80% complete): by looking at a setting where large amounts of context can be gleaned in an unsupervised manner and combined with translation information from parallel corpora, we test our low-dimensional hypothesis in extremely high-dimensional scenarios. Specifically, we consider a nearest-neighbor approach to translation where a phrase pair is embedded in a graph, the source side a node with edges to its k most similar phrases, and the target side as a "label" for the node. The presence and strengths of these edges is determined by the context in which the phrases occur, and is computed on monolingual corpora. Translation information can propagate through the graph either enabling the discovery of new phrases and their translations, or in a domain adaptation setting where we use in-domain monolingual data to embed translation units. The framework allows an interesting empirical evaluation of context dimensionality in an MT setting, since the graphs can be constructed in a number of different ways, either in the raw high-dimensional space or in a recovered low-dimensional one. Furthermore, in order to learn context-based representations of higher-order n-grams existing methods scale poorly, which introduces the need for learning compositional functions to construct these graphs, and which are evaluated in terms of impact on downstream MT performance.

2 Latent-Variable Models for Structured Context

The first component of this thesis focuses on low-rank embeddings of translation rules of the kind found in hierarchical and syntax-based MT systems, and is formulated as a data-driven refinement of SCFGs with latent, syntactic categories for non-terminals (NTs), without the use of additional resources and completely from the parallel data. Previous approaches in MT have attempted this problem either with external resources e.g., a source language parser (Zollmann and Venugopal, 2006, Huang et al., 2010), or in conjunction with synchronous grammar induction (Blunsom et al., 2008a, Mylonakis and Sima'an, 2011), while we assume a fixed grammar *a priori*, and work with a *minimal grammar*, extracted by considering the smallest set of rules that explains each parallel sentence given a word alignment (Zhang et al., 2008). We refine this grammar in a context-dependent manner by introducing non-terminal distinctions to capture contextual regularities. Hence, a secondary aim of this section is to

evaluate translation performance with these minimal grammars in a setup where context dependence is incorporated low-dimensionally and not through composed grammars.

		BLEU	
	Setup	Dev	Test
Baselines	HIERO	46.08	55.31
	MIN-GRAMMAR	43.38	51.78
	MLE	43.24	52.80
Spectral	$m = 1$ RI	44.18	52.62
	$m = 8$ RI	44.60	53.63
	$m = 16$ RI	46.06	**55.83**
	m=16 RI+Lex+Sm	**46.08**	55.22
	m=16 RI+Lex+Len	45.70	55.29
	m=24 RI+Lex	43.00	51.28
	m=32 RI+Lex	43.06	52.16
EM	$m = 8$	40.53 (0.2)	49.78 (0.5)
	$m = 16$	42.85 (0.2)	52.93 (0.9)
	$m = 32$	41.07 (0.4)	49.95 (0.7)

Table 1: Results for the ZH-EN corpus, comparing across the baselines and the two parameter estimation techniques. RI, Lex, and Len correspond to the rule indicator, lexical, and length features respectively, and Sm denotes smoothing. For the EM experiments, we selected the best scoring iteration by tuning weights for parameters obtained after 25 iterations and evaluating other parameters with these weights. Results for EM are averaged over 5 starting points, with standard deviation given in parentheses. Spectral, EM, and MLE performances compared to the MIN-GRAMMAR baseline are statistically significant ($p < 0.01$).

We compare two ways of learning the grammar refinement: a likelihood-maximization approach using expectation maximization (EM), and a moments-based approach using SVD, which is a generalization of the monolingual version in Cohen et al. (2012) to the translation setting. In the latter, the latent m-dimensional space is recovered by computing a truncated SVD of a feature covariance matrix, where the features are associated with inside and outside sub-trees of NTs that represent instantiations of translation rules. We show using experiments on MT that estimating parameters in the SVD-based space is empirically more effective than using EM and also better than working in a high-dimensional space. Table 1 presents results using the BTEC Chinese-English corpus (Paul, 2009), which highlights the relative performance of the two parameter estimation techniques. In fact, the minimal grammar and SVD-estimated latent-variable model combination matches the performance of a more conventional "hiero" grammar (Chiang, 2007). For both the EM and spectral estimation, we experimented with several different latent space sizes and feature configurations.

The work in this section is largely complete and published in Saluja et al. (2014a). For future work in this thesis, we propose the following extensions:

- better use of the refined grammar at decoding time (e.g., exploring the effects of including inside and outside probabilities for each latent state instead of or in combination with the rule marginals).
- better features in the spectral algorithm (e.g., lexical and phrasal relative frequency estimates), including the usage of learned, distributed representations (Mikolov et al.,

2013b) instead of (or in combination with) sparse "one-hot" representations for lexical features. Previous experiments where words are replaced with their Brown cluster IDs were inconclusive.

The contributions are: a generalization of latent PCFGs (Matsuzaki et al., 2005) to latent SCFGs; an efficient tensor-based version of the inside-outside algorithm; an empirical demonstration that adding marginal rule probabilities from this model as features in the traditional linear translation model (Och and Ney, 2004) improves translation quality; and two algorithms that learn these latent categories (equivalently, the latent space) from the data without any externally imposed syntactic labels.

3 Multiple Views for Low-Dimensional Context

In most scenarios during test time, the MT decoder has access to an entire source document prior to translation; it is thus absurd that translation models use very little, if any, of the large amounts of observable context. Hence, the second portion of this thesis concentrates on modeling the source context *directly* in a low-dimensional space.[3] As in §2, we use minimal grammars instead of composed ones and shift the context dependence to the lower-dimensional space, which makes estimation and inference more tractable. Recent work has shown that extreme source-side context can be leveraged to great effect to improve translation (Devlin et al., 2014), as long as the representation of this context is manageable, which the authors achieve through neural network-based representations. While neural networks are amazingly expressive, they are notoriously difficult to train (Pascanu et al., 2013, *inter alia*), and instead we propose a simple method to recover a linear low-dimensional subspace which leverages the multi-view assumption stated before, where we use multiple "views" of the data to learn an appropriate low-dimensional basis in order to manage extremely large amounts of context.

The basic setting we pursue is reminiscent of "word-sense disambiguation" for MT (Carpuat, 2008), in that we condition on source contextual information to select the appropriate translation rule (e.g., source-target phrase pair) to be used in that context. However, by learning an appropriate low-dimensional basis first, we avoid significant feature engineering, and can take advantage of the reduced sample complexities. In order to learn such a basis for representing context and rules, we make use of CCA (Hardoon et al., 2004, Kakade and Foster, 2007). For occurrences of translation units in a parallel corpus, the source-side context of the rule can be split into multiple views: a natural one is context that occurs before the rule and context that occurs after, although alternate splits, e.g., lexical-based features vs. class or POS-based features, or inside vs. outside tree features, will also be explored. Global features, namely those that exist at the paragraph or document-level, can also be considered in this setting. Then, CCA is performed between a matrix corresponding to context tokens and a matrix corresponding to phrase pair tokens, which yields a latent space where the two sets of random variables (corresponding to context and phrase pairs) are maximally correlated.

[3] Including target context breaks one of the key independence assumptions made by phrase-based translation models, that translations of source phrases are conditionally independent of each other, given the source sentence. Target context is unobserved during evaluation, so conditioning upon this information directly in the translation model is computationally more difficult. Outside of the translation model, the language model also takes care of target-side context dependencies.

CCA recovers a pair of projection matrices, one for the context (i.e., projects a sparse context to its dense low-dimensional representation) and one for the translation units, from which we can reconstruct the representations for both in the shared latent space.

For evaluation, as in §2 we once again ask if we can achieve effective translation with minimal grammars, but this time concentrating our efforts on the context representations. We also investigate if translation can be improved by considering global context. Our context representations can be incorporated into MT in a number of different ways. We can add the translation unit representations (which are low-dimensional representations of context) as additional features in an MT tuning setup, as proposed for the latent state representations in §2, and tune their weights directly. Alternatively, we can use the recovered space as a basis for supervised learning. For example, the context projection matrix acquired during training can be used to recover latent representations of context during test time. With a small number of similarity computations (restricted to the translation rules for a given source phrase), we can recover the translation units that are closest to the context in the low-dimensional space through a k-NN setup. Or, weights for a linear model that scores various translation options available to a source phrase conditioned directly on context can be learned with least-squares estimation (Foster et al., 2008).

One can argue that the previous setting does not fully utilize the supervision available; context is estimated over a parallel corpus to understand how translation units translate, but the representations do not explicitly take into account label information. To this end, we can utilize a two-step regression approach (Shah et al., 2010). As before, the approach splits the training data; with one view, we approximate the least-squares solution by assuming no covariance among features. This view is subsequently refined on a development set by using shrinkage estimators (a form of regularization) to fine-tune each of the predictors. The first step consisting of the diagonal approximation can be seen as an example of discriminative representation learning, where we accumulate the co-occurrences of context-based features with their "labels" (i.e., translations); these features are then tuned to optimize a translation task through shrinkage estimators. In this approach, we jointly learn to classify over all target translations in a scalable manner instead of independent classifications for each source phrase. It would be interesting to further explore the links between the CCA-based approaches, where the label information is used in conjunction with the unlabeled information to recover an appropriate latent space, and the shrinkage-based approach.

The contributions are: two principled approaches to translation sense disambiguation that work directly with low-dimensional context representations; and empirical evidence that shows incorporating extreme local context as well as global context in conjunction with minimal grammars achieves performance equivalent to or better than large, composed grammars.

4 Low-Dimensional Context & Semi-Supervised Learning

Importantly, the source language context information we seek is not restricted to parallel sentence corpora and representations can be learned from the much more copious amount of monolingual data available. In this section, we investigate the central hypothesis of this thesis in the semi-supervised learning (SSL) setting. A semi-supervised approach allows us to test our hypothesis in the limit: by extracting context over large monolingual corpora, we

have access to rich, variable contexts that exist in an extremely high-dimensional space of size proportional to vocabulary size. Since we hypothesize that the salient context information on which our models should condition resides in a low-dimensional space, this perspective should do better than the high-dimensional view, keeping in mind the obvious computational benefits it also introduces.

In particular, we adopt a graph-based SSL setup (with phrases constituting nodes in a k-nearest neighbor graph) and perform inference over the graph via label propagation (Zhu, 2005). Using contextual similarity to construct graphs is an empirically effective approach and widely adopted in NLP (Subramanya et al., 2010, Das and Petrov, 2011, *inter alia*), but more importantly from our perspective using low-dimensional representations of this context prior to graph construction reduces the intrinsic dimensionality of the problem. The graphs are discrete approximations of continuous manifolds, which form non-linear subspaces of the original high-dimensional space. Therefore, recovering low-dimensional representations of phrases prior to embedding them in the graph means that the training data is a more dense sampling of the manifold structure, resulting in sample complexity benefits too. In light of these observations, it is surprising that little previous work evaluating the various low-dimensional representations and how they improve graph quality and performance in various tasks exists, and none in MT. This section of the thesis aims to address these gaps.

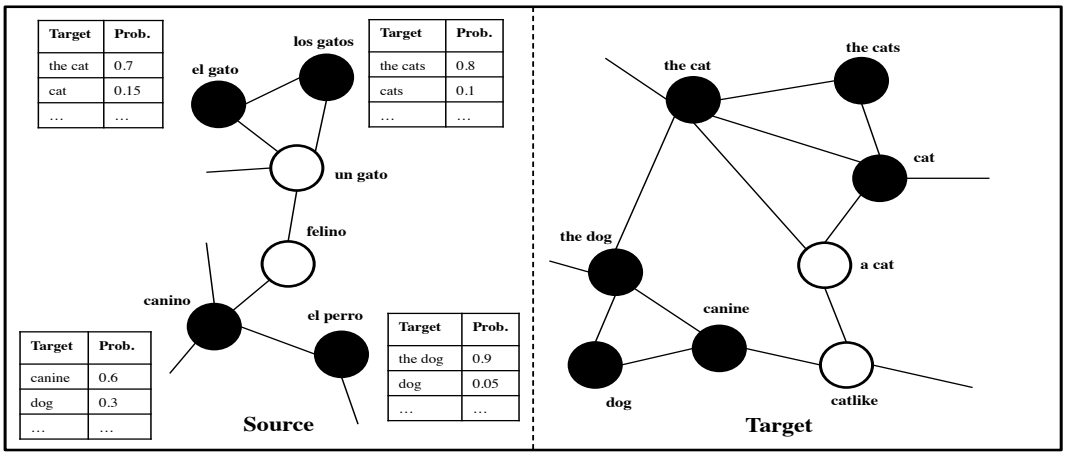

Figure 1: Example source and target graphs used in our approach. Labeled phrases on the source side are black (with their corresponding translations on the target side also black); unlabeled and generated phrases on the source and target sides respectively are white. Labeled phrases also have conditional probability distributions defined over target phrases, which are extracted from the parallel corpora.

Monolingual data is used to construct the similarity graphs over phrases (word sequences), with separate graphs for the source and target languages (Fig. 1). The source similarity graph consists of phrase nodes representing sequences of words in the source language. If a source phrase is found in the baseline phrase table it is called a **labeled** phrase: its conditional relative frequency estimates over target phrases (estimated from the parallel data) is used as the label, and is subsequently never changed. Otherwise it is called an **unlabeled** phrase, and our algorithm finds labels (translations) for these unlabeled phrases, with the help of the graph-based representation. The label space is thus the phrasal translation inventory, and like the source side it can also be represented in terms of a graph. We take the additional step of enriching the target graph by generating additional candidate translations by auxiliary

	BLEU	
Setup	**Tune**	**Test**
Baseline	39.33	38.09
SLP 1-gram	39.47	37.85
LP 2-gram	40.75	38.68
SLP 2-gram	41.00	39.22
SLP-HalfMono 2-gram	40.82	38.65
SLP+Morph 2-gram	41.02	39.35
Baseline+LargeLM	41.48	39.86
SLP+LargeLM	42.82	41.29

Table 2: Results for the Arabic-English evaluation. The LP vs. SLP comparison highlights the importance of target side enrichment via translation candidate generation, 1-gram vs. 2-gram comparisons highlight the importance of emphasizing phrases, utilizing half the monolingual data shows sensitivity to monolingual corpus size, and adding morphological information results in additional improvement.

means, e.g., through the baseline translation system as in self-training (McClosky et al., 2006) or through morphological analyzers (Toutanova et al., 2008).

Downstream, we leverage the graph structure to suggest and score translations for previously unseen source language phrases by propagating information from source phrases present in the parallel corpus to similar unlabeled phrases. Results for an Arabic-English setup (17M training tokens), evaluated on standard NIST corpora (MT06 and MT08) are presented in Table 2. We explore two graph propagation algorithms, the standard label propagation (Zhu and Ghahramani, 2002) as well as a structured variant (Liu et al., 2012) that also takes into account the target graph structure when scoring these translations. The large language model experiments utilize a language model trained on 7.6 billion tokens, and here the improvement over the baseline is even larger. Similar results were also obtained for Urdu-English (Saluja et al., 2014b).

The graph-based SSL framework in this section is largely complete. For future work in this thesis, we propose the following extensions:

- use lower-dimensional contextual representations of phrases via SVD, CCA using the multi-view assumption, and using learned, distributed representations to improve phrasal similarity.
- re-purpose the approach for domain adaptation. Initial translation distributions estimated from out-of-domain data can be re-estimated based on a small amount of in-domain parallel data and a large amount of in-domain monolingual data. Contextual information from the new domain is thus reflected in the graph structure.
- improve scalability. Graph construction is the current computational bottleneck, both in terms of feature extraction for longer translation units like phrases, and the quadratic dependence of explicit graph construction on the number of translation units (which increases exponentially in the length of the unit). These computational limitations encourage a move towards the *continuous* version of the k-NN paradigm, where the phrasal representations can be constructed dynamically. In §4.1 we discuss future work along these lines where a model for compositional vector-space semantics that directly takes into account non-compositionality is proposed, applied towards improving graph

embeddings for phrasal units.

The contributions of this section are: a graph-based SSL method to expand translation models with information contained in monolingual data via contextual similarity; modification of this method to adapt existing translation distributions in a domain adaptation setting; and an evaluation of various dimensionality reduction techniques in their approximation of a non-linear manifold as represented by the graph.

4.1 Compositional Functions for Phrasal Representations

Graph construction depends on vector representations of phrases, and vector space models for words have risen in popularity recently, along with compositional models that combine word representations to infer the semantics of longer units like phrases or sentences (Mitchell and Lapata, 2010, Socher et al., 2012, *inter alia*). These compositional models attempt to circumvent the obvious computational and statistical issues that arise from estimating these multiword representations directly: there are far more units that need to be handled and the occurrences of such units in training corpora diminish rapidly. This line of work has been interesting, but mostly unproven on downstream applications. We suggest MT is an excellent testbed for applying compositional models, but the issue of detecting non-compositional phrases i.e., multi-word expressions whose meaning cannot be inferred from or is not a function of the semantics of its constituents (and thus, combining constituents in a compositional manner would yield the incorrect meaning), becomes paramount. Non-compositional phrases e.g., *hot line* are translated with often hilarious (or disastrous) consequences, as by definition these phrases translate idiosyncratically.

We propose to model compositionality using bilinear maps (tensors) and learn the parameters through a principled regression-based framework (similar to Grefenstette et al. (2013)). However, our functions apply at the level of part-of-speech categories and are not lexicalized, avoiding an overuse of parameters. Heuristic priors on parameter structure are not imposed, and instead we leave feature and structure recovery to the learning algorithms. To provide training data for our functions, we make use of the Paraphrase Database (Ganitkevitch et al., 2013, PPDB), which contains one-to-many and many-to-one paraphrases (monolingual translations) to extract training examples for each compositional relationship (by definition, we cannot learn non-compositional representations from constituents). PPDB contains examples such as "staff member $\leftrightarrow$ official"; thus, we assume the vectors for "staff" and "member" combine compositionally and map to the vector for "official". The problem then boils down to multivariate multiple regression, where the independent variable is one component of the output vector and the dependent variables are the pairwise interactions of components across the two vectors. Preliminary results on standard evaluation corpora in this task that compare phrasal similarities computed with composed representations with human judgments (Mitchell and Lapata, 2010) yield state-of-the-art results for adjective-noun and noun-noun compositions.

For non-compositionality detection, the basic intuition we leverage is that a phrasal representation should be predictive of the actual context the phrase occurs in during test time. Features based on the likelihood of the context given the phrasal representation (similar to the objective function in the skip-gram model of Mikolov et al. (2013a)), along with the context and phrase representations themselves, are used to model the latent phrasal segmentation of

a source sentence through a discriminative latent-variable model (Blunsom et al., 2008b). In this manner, we can reason about the importance of non-compositionality with features based on the likelihood of context when learning an optimal phrasal segmentation that explains the parallel data.

The contributions of this section are: a scalable, data-driven method that uses paraphrase information to learn compositional functions; an evaluation of compositional functions on a new downstream task; a scheme that leverages the compositional functions to detect non-compositional expressions in a corpus; and empirical evidence that non-compositional knowledge matters in MT.

5 Tentative Timeline

Large parts of the thesis have already been completed. The following timeline outlines the order in which remaining work will be tackled, as well as the expected amount of time each task will take. Since internal deadlines have been aligned with external deadlines (conference submissions) and the basic approaches have already been constructed, the timeline is reasonable in its expectations.

- October - December 2014: Complete non-compositionality detection and experiments with MT (§4.1); submit to NAACL 2015 (December 4, 2014 deadline).
- December 2014 - February 2015: implementation of and experimentation with multi-view context representations (§3); submit to ACL 2015 (February 27, 2014 deadline).
- March - April 2015: extensions to latent-variable models for structured context (§2) and semi-supervised models for MT (§4).
- May - July 2015: thesis writing.
- August 2015: thesis defense.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March 1996. ISSN 0891-2017.

Phil Blunsom, Trevor Cohn, and Miles Osborne. Bayesian Synchronous Grammar Induction. In *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, NIPS 2008, 2008a.

Phil Blunsom, Trevor Cohn, and Miles Osborne. A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 200–208, Columbus, Ohio, June 2008b.

Marine Carpuat. *Word Sense Disambiguation for Statistical Machine Translation*. PhD thesis, Hong Kong University of Science and Technology, Hong Kong, 2008.

Marine Carpuat, Hal Daume III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

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

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393, 1999.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June 2007.

S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. Ungar. Spectral learning of latent-variable PCFGs. In *Proceedings of ACL*, 2012.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL*, 2011.

Hal Daumé III and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Association for Computational Linguistics*, Portland, OR, 2011.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Multi-View Learning of Word Embeddings via CCA. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, 2011.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Yang Feng and Trevor Cohn. A markov model of machine translation using non-parametric bayesian inference. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 333–342, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Dean P. Foster, Sham M. Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. Technical report, Toyota Technological Institute (TTI), 2008.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 961–968. Association for Computational Linguistics, 2006.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Kevin Gimpel and Noah A. Smith. Rich source-side context for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 9–17, 2008.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. 2013.

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, December 2004. ISSN 0899-7667.

Zhongjun He, Qun Liu, and Shouxun Lin. Improving statistical machine translation using lexicalized rule selection. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 321–328, 2008.

Liang Huang, Kevin Knight, and Aravind Joshi. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA*, August 2006.

Zhongqiang Huang, Martin Čmejrek, and Bowen Zhou. Soft syntactic constraints for hierarchical phrase-based translation using latent syntactic distributions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 138–147. Association for Computational Linguistics, 2010.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In *Proceedings of the 20th Annual Conference on Learning Theory*, COLT'07, pages 82–96, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for*

Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Shujie Liu, Chi-Ho Li, Mu Li, and Ming Zhou. Learning translation consensus with structured label propagation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 302–310, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. N-gram-based machine translation. *Comput. Linguist.*, 32(4):527–549, December 2006. ISSN 0891-2017.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Probabilistic cfg with latent annotations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 75–82. Association for Computational Linguistics, 2005.

David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, June 2006. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013b. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439, 2010.

Markos Mylonakis and Khalil Sima'an. Learning Hierarchical Translation Structure with Linguistic Annotations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 642–652, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pages 160–167, July 2003.

Franz Josef Och and Hermann Ney. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449, December 2004.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the International Conference of Machine Learning*, Atlanta, Georgia, June 2013.

Michael Paul. Overview of the iwslt 2009 evaluation campaign. In *Proceedings of IWSLT 2009*, 2009.

Avneesh Saluja, Chris Dyer, and Shay B. Cohen. Latent-Variable Synchronous CFGs for Hierarchical Translation. In *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing, EMNLP '14. Association for Computational Linguistics, 2014a.

Avneesh Saluja, Kristina Toutanova, Chris Quirk, and Hany Hassan. Graph-based semi-supervised learning of translation models from monolingual data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL-52. Association for Computational Linguistics, 2014b.

Rushin Shah, S. Paramveer Dhillon, Mark Liberman, Dean Foster, Mohamed Maamouri, and Lyle Ungar. A new approach to lexical disambiguation of arabic text. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 725–735. Association for Computational Linguistics, 2010.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012.

N. Stroppa, A. Van den Bosch, and A. Way. Exploiting source similarity for SMT using context-informed features. In A. Way and B. Gawronska, editors, *Proceedings of the 11th International Conference on Theoretical Issues in Machine Translation (TMI 2007)*, 2007.

Amarnag Subramanya, Slav Petrov, and Fernando Pereira. Efficient graph-based semi-supervised learning of structured tagging models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 167–176, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Ashish Vaswani, Haitao Mi, Liang Huang, and David Chiang. Rule markov models for fast tree-to-string translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 856–864, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Hao Zhang, Daniel Gildea, and David Chiang. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 1081–1088. Association for Computational Linguistics, 2008.

Xiaojin Zhu. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005. AAI3179046.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, 2002.

Andreas Zollmann and Ashish Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, pages 138–141, 2006.