Composed, Distributed Reflections on Semantics and Statistical Machine Translation

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Composed, Distributed Reflections on Semantics and Statistical Machine Translation ... A Hitchhiker's Guide

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

1. Elements of a Compositional, Distributed SMT Model
2. Training a Compositional, Distributed SMT Model
3. Semantics and SMT
4. Moving Forward
5. Summary

The Nature of a Word Representation I

- **Distributed representation:** words are projected into an n-dimensional real-valued space with "dense" values [Hinton et al., 1986]

$$
\begin{array}{rl}
\textit{bicyle}: & [\ 0.834 \quad -0.342 \quad 0.651 \quad 0.152 \quad -0.941\ ] \\
\textit{cycling}: & [\ 0.889 \quad -0.341 \quad -0.121 \quad 0.162 \quad -0.834\ ]
\end{array}
$$

- **Local representation:** words are projected into an n-dimensional real-valued space using a "local"/one-hot representation:

$$
\begin{array}{rccccccc}
 & & & \textit{bicycle} & & \textit{cycling} & & \\
\textit{bicycle}: & [& \ldots & 1 & \ldots & 0 & \ldots &] \\
\textit{cycling} & [& \ldots & 0 & \ldots & 1 & \ldots &]
\end{array}
$$

The Nature of a Word Representation II

- In the multilingual case, ideally project words from different languages into a common distributed space:

$$
\begin{aligned}
\textit{bicycle}_{\texttt{EN}} &: \; [\; 0.834 \quad -0.342 \quad 0.651 \quad 0.152 \quad -0.941 \;] \\
\textit{cycling}_{\texttt{EN}} &: \; [\; 0.889 \quad -0.341 \quad -0.121 \quad 0.162 \quad -0.834 \;] \\
\textit{Rad}_{\texttt{DE}} &: \; [\; 0.812 \quad -0.328 \quad -0.113 \quad 0.182 \quad -0.712 \;] \\
\textit{Radfahren}_{\texttt{DE}} &: \; [\; 0.832 \quad -0.302 \quad 0.534 \quad 0.178 \quad -0.902 \;]
\end{aligned}
$$

The Basis of a Word Representation I

- **Representational basis:** the basis of the projection for word $w \in V$ is generally some form of "distributional" model, conventionally in the form of some aggregated representation across token occurrences w_i of "contexts of use" $\mathrm{ctxt}(w_i)$:

$$\mathrm{dsem}(w) = \mathrm{agg}(\{\mathrm{ctxt}(w_i)\})$$

The Basis of a Word Representation II

- "Context of use" represented in various ways, incl. bag-of-words, positional words, bag-of-*n*-grams, and typed syntactic dependencies [Pereira et al., 1993, Weeds et al., 2004, Padó and Lapata, 2007]

... to ride a	*bicycle*	*or solve puzzles ...*
... produced a heavy-duty	*bicycle*	*tire that outlasted ...*
... now produces 1,000	*bicycle*	*and motorbike tires ...*
... Peterson mounts her	*bicycle*	*and grinds up ...*
... some Marin County	*bicycle*	*enthusiasts created a ...*

- **First-order model** = context units represented "directly"; **second-order models** = context represented via distributional representation of each unit; ...

Compositional Semantics

- **Compositional semantic model** = model the semantics of an arbitrary combination of elements (p) by composing together compositional semantic representations of its component elements ($p = \langle p_1, p_2, ...\rangle$); for "atomic" elements, model the semantics via a distributed (or otherwise) representation:

$$\mathrm{csem}(p) = \begin{cases} \mathrm{dsem}(p) & \text{if } p \in V \\ \mathrm{csem}(p_1) \circ \mathrm{csem}(p_2) ... & \text{otherwise} \end{cases}$$

Source(s): Mitchell and Lapata [2010]

Comparing Representations

- For both word and compositional semantic representations, "comparison" of representations is generally with simple cosine similarity, or in the case of probability distributions, scalar product, Jensen-Shannon divergence, or similar

Source(s): Dinu and Lapata [2010], Lui et al. [2012]

Talk Outline

1. Elements of a Compositional, Distributed SMT Model
2. Training a Compositional, Distributed SMT Model
3. Semantics and SMT
4. Moving Forward
5. Summary

Learning Word Representations I

- Two general approaches [Baroni et al., 2014]:
  1. **Count**: count up word co-occurrences in context window of some size, across all occurrences of a given target word; generally perform some smoothing, weighting and dimensionality reduction over this representation to produce a distributed representation
  2. **Predict**: use some notion of context similarity and discriminative training to learn a representation whereby the actual target word has better fit with its different usages, than some alternative word [Collobert et al., 2011]

Learning Word Representations II

- In the immortally-jaded words of [Baroni et al., 2014, p244–245]:

> *As seasoned distributional semanticists ... we were annoyed by the triumphalist overtones often surrounding predict models ... Our secret wish was to discover that it is all hype, and count vectors are far superior to their predictive counterparts. A more realistic expectation was that a complex picture would emerge ... Instead, we found that the predict models are so good that, while the triumphalist overtones still sound excessive, there are very good reasons to switch to the new architecture.*

Sample Count Methods

- **Term weighting:** positive PMI, log-likelihood ratio

Sample Count Methods

- **Term weighting:** positive PMI, log-likelihood ratio
- **Dimensionality reduction:** SVD, non-negative matrix factorisation

Sample Count Methods

- **Term weighting:** positive PMI, log-likelihood ratio
- **Dimensionality reduction:** SVD, non-negative matrix factorisation
- **"Standalone" methods:**
  - **Brown clustering [Brown et al., 1992]:** hierarchical clustering of words based on maximisation of bigram mutual information
  - **Latent Dirichlet allocation (LDA: Blei et al. [2003]):** construct term–document matrix (possibly with frequency-pruning of terms), and learn T latent "topics" (term multinomials per topic) and topic allocations (topic multinomials per document); derive word representations via the topic allocations across all usages of a target word

Approaches to Composition

- Two general approaches:
  1. Apply a predefined operator to the component (vector) representations, e.g. (weighted) vector addition, matrix multiplication, tensor product, ... [Mitchell and Lapata, 2010]
  2. (Hierarchically) learn a composition weight matrix, and apply a non-linear transform to it at each point of composition [Mikolov et al., 2010, Socher et al., 2011, 2012, Mikolov et al., 2013]

Sample Learned Compositional Methods

- **Recursive neural networks [Socher et al., 2012, 2013]):** jointly learn composition weight vector(s) and tune word embeddings in a non-linear bottom-up (binary) recursive manner from the components
  - optional extras: multi-prototype word embeddings [Huang et al., 2012], incorporation of morphological structure [Luong et al., 2013]
- **Recurrent neural networks [Mikolov et al., 2010, 2013]:** learn word embeddings in a non-linear recurrent manner from the context of occurrence

Talk Outline

1. Elements of a Compositional, Distributed SMT Model
2. Training a Compositional, Distributed SMT Model
3. Semantics and SMT
4. Moving Forward
5. Summary

Semantics and MT: pre/ex-SMT

- Back in the day of RBMT, (symbolic) lexical semantics was often front and centre (esp. for distant language pairs), including:
  - interlingua [Mitamura et al., 1991, Dorr, 1992/3]
  - formal lexical semantics [Dorr, 1997]
  - verb classes and semantic hierarchies used for disambiguation/translation selection and discourse analysis [Knight and Luk, 1994, Ikehara et al., 1997, Nakaiwa et al., 1995, Bond, 2005]

Semantics and MT: pre/ex-SMT

- Back in the day of RBMT, (symbolic) lexical semantics was often front and centre (esp. for distant language pairs), including:
  - interlingua [Mitamura et al., 1991, Dorr, 1992/3]
  - formal lexical semantics [Dorr, 1997]
  - verb classes and semantic hierarchies used for disambiguation/translation selection and discourse analysis [Knight and Luk, 1994, Ikehara et al., 1997, Nakaiwa et al., 1995, Bond, 2005]
- There is also an ongoing traditional of work on compositional (formal) semantics in MT, based on deep parsing [Bojar and Hajič, 2008, Bond et al., 2011]

Semantics and MT: Enter SMT I

- In the space of SMT, many have attempted to make use of (lexical) semantics, but few success stories, notably:
  - Vickrey et al. [2005]: WSD-based models enhance "word translation" (fill-in-the-blank MT)
  - Cabezas and Resnik [2005]: source "word senses" via word alignment, and train a WSD system over them; inject translations into the phrase table based on the (soft) predictions of the WSD model
  - Chan et al. [2007]: WSD-style disambiguation model predictions incorporated into Hiero improve SMT
  - Carpuat and Wu [2007]: integrating WSD-style models into the SMT decoder and disambiguating over phrasal translation candidates improves SMT

Semantics and MT: Enter SMT II

- Zhao and Xing [2007], Xiao et al. [2012], Eidelman et al. [2012]: biasing the translation model with topic model-based features improves SMT

Semantics and MT: Enter SMT II

- Zhao and Xing [2007], Xiao et al. [2012], Eidelman et al. [2012]: biasing the translation model with topic model-based features improves SMT
- Carpuat et al. [2013]: when moving to new domains, incorporation of "new sense" information into the phrase table improves SMT

Semantics and MT: Enter SMT II

- Zhao and Xing [2007], Xiao et al. [2012], Eidelman et al. [2012]: biasing the translation model with topic model-based features improves SMT
- Carpuat et al. [2013]: when moving to new domains, incorporation of "new sense" information into the phrase table improves SMT

- Instances of methods which successfully use an explicit representation of word sense are much harder to find:
  - Xiong and Zhang [2014]: improvements in SMT through: (1) performing all-words WSI based on topic modelling [Lau et al., 2012]; (2) training per-word disambiguation models conditioned on the sense assignment; and (3) incorporation of the translation predictions into the decoder

Semantics and MT Evaluation

- More joy in the MT evaluation metric space, e.g.:
  - Liu et al. [2010], Dahlmeier et al. [2011]: the inclusion of WordNet-based synonym features into TESLA improves the metric
  - Denkowski and Lavie [2011]: the inclusion of WordNet synset overlap into the unigram matching component of METEOR improves the metric
  - Lo and Wu [2011], Lo et al. [2012]: MT evaluation based on shallow semantic parsing + automatic semantic frame alignment correlates better than string-based methods for adequacy-based evaluation

Semantics and Multilingual Text

- Numerous examples of multilingual text improving semantic analysis, including:
  - WSD [Dagan and Itai, 1994, Diab and Resnik, 2002, Ng et al., 2003, Tufiș et al., 2004]
  - paraphrase detection [Barzilay and McKeown, 2001, Dolan et al., 2004, Bannard and Callison-Burch, 2005]
  - PP attachment disambiguation [Schwartz et al., 2003]
  - wordnet construction [Bentivogli and Pianta, 2005, Bond et al., 2012]

Outline of Possible Extra Components in a Neural SMT System

- Neural language model jointly conditioned on the target and source languages [Le et al., 2012, Kalchbrenner and Blunsom, 2013, Devlin et al., 2014]
- Bilingual word embeddings [Zou et al., 2013]
- Dynamic pooling [Socher et al., 2011] or convolutional NNs [Kalchbrenner and Blunsom, 2013] to capture (pseudo-)syntax

Neural SMT

- After years of attacking a very solid brick wall with different semantic battering rams, neural SMT models are producing big gains in MT accuracy across the board ... what gives?

Neural SMT

- After years of attacking a very solid brick wall with different semantic battering rams, neural SMT models are producing big gains in MT accuracy across the board ... what gives?
  - it is well established that neural LMs are more accurate than conventional LMs [Mikolov et al., 2010]

Neural SMT

- After years of attacking a very solid brick wall with different semantic battering rams, neural SMT models are producing big gains in MT accuracy across the board ... what gives?
  - it is well established that neural LMs are more accurate than conventional LMs [Mikolov et al., 2010]
  - neural LMs are also more expressive, opening up the possibility of jointly modelling the source and target language strings [Le et al., 2012, Kalchbrenner and Blunsom, 2013, Devlin et al., 2014]

Neural SMT

- After years of attacking a very solid brick wall with different semantic battering rams, neural SMT models are producing big gains in MT accuracy across the board ... what gives?
  - it is well established that neural LMs are more accurate than conventional LMs [Mikolov et al., 2010]
  - neural LMs are also more expressive, opening up the possibility of jointly modelling the source and target language strings [Le et al., 2012, Kalchbrenner and Blunsom, 2013, Devlin et al., 2014]
  - convolutional NNs et al. appear to be an effective means of "continuous" syntactic and semantic composition

Neural SMT

- After years of attacking a very solid brick wall with different semantic battering rams, neural SMT models are producing big gains in MT accuracy across the board ... what gives?
  - it is well established that neural LMs are more accurate than conventional LMs [Mikolov et al., 2010]
  - neural LMs are also more expressive, opening up the possibility of jointly modelling the source and target language strings [Le et al., 2012, Kalchbrenner and Blunsom, 2013, Devlin et al., 2014]
  - convolutional NNs et al. appear to be an effective means of "continuous" syntactic and semantic composition
  - ... more to the point, what is semantics anyways?

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...
    - are the new-generation neural SMT models really semantic?

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...
  - are the new-generation neural SMT models really semantic?
  - NNLMs vs. sense-based translation partitioning vs. better context modelling ... aren't we comparing apples and oranges?

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...
  - are the new-generation neural SMT models really semantic?
  - NNLMs vs. sense-based translation partitioning vs. better context modelling ... aren't we comparing apples and oranges?
  - the difference between formal compositional models and distributed models may be less than it would appear [Grefenstette, 2013, Beltagy et al., 2013, Lewis and Steedman, 2013]

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...
  - are the new-generation neural SMT models really semantic?
  - NNLMs vs. sense-based translation partitioning vs. better context modelling ... aren't we comparing apples and oranges?
  - the difference between formal compositional models and distributed models may be less than it would appear [Grefenstette, 2013, Beltagy et al., 2013, Lewis and Steedman, 2013]
  - what is semantics anyway?

    Semantics ... focuses on the relation between signifiers ... and what they stand for, their denotation. (Wikipedia 25/10/14)

Semantics and SMT

- Where semantics had largely failed to deliver in the past, it is now seemingly delivering wholesale ...
  - are the new-generation neural SMT models really semantic?
  - NNLMs vs. sense-based translation partitioning vs. better context modelling ... aren't we comparing apples and oranges?
  - the difference between formal compositional models and distributed models may be less than it would appear [Grefenstette, 2013, Beltagy et al., 2013, Lewis and Steedman, 2013]
  - what is semantics anyway?

    Semantics ... focuses on the relation between signifiers ... and what they stand for, their denotation. (Wikipedia 25/10/14)
  - Important to bear in mind that the storyline in other areas of NLP is strikingly similar

Talk Outline

1. Elements of a Compositional, Distributed SMT Model
2. Training a Compositional, Distributed SMT Model
3. Semantics and SMT
4. Moving Forward
5. Summary

Lexical Semantic Approaches via Distributed, Compositional Approaches

- Returning to the isolated success stories of semantics improving SMT, what were they doing right, and what are the implications for neural SMT?

Lexical Semantic Approaches via Distributed, Compositional Approaches

- Returning to the isolated success stories of semantics improving SMT, what were they doing right, and what are the implications for neural SMT?
- In the case of Cabezas and Resnik [2005], Chan et al. [2007] and Carpuat and Wu [2007], I would argue that the success of the model came from richer representations/disambiguation of context embedded in an SMT context ... which is the greatest advantage offered by distributed approaches to SMT

Lexical Semantic Approaches via Distributed, Compositional Approaches

- Returning to the isolated success stories of semantics improving SMT, what were they doing right, and what are the implications for neural SMT?
- In the case of Cabezas and Resnik [2005], Chan et al. [2007] and Carpuat and Wu [2007], I would argue that the success of the model came from richer representations/disambiguation of context embedded in an SMT context ... which is the greatest advantage offered by distributed approaches to SMT
- In the case of Carpuat et al. [2013], pre-training over (target language) data in the novel domain can potentially capture the necessary mapping onto the "old" domain to substitute for the domain dictionary etc.

Distributed, Compositional Models and MT Evaluation

- Clear scope to incorporate (general-purpose) distributed word/phrase representations into MT evaluation metrics

Distributed, Compositional Models and MT Evaluation

- Clear scope to incorporate (general-purpose) distributed word/phrase representations into MT evaluation metrics
- Also possibility to include composition into models (a la Socher et al. [2011] or Kalchbrenner and Blunsom [2013])

Distributed, Compositional Models and MT Evaluation

- Clear scope to incorporate (general-purpose) distributed word/phrase representations into MT evaluation metrics
- Also possibility to include composition into models (a la Socher et al. [2011] or Kalchbrenner and Blunsom [2013])
- Words of caution:

Distributed, Compositional Models and MT Evaluation

- Clear scope to incorporate (general-purpose) distributed word/phrase representations into MT evaluation metrics
- Also possibility to include composition into models (a la Socher et al. [2011] or Kalchbrenner and Blunsom [2013])
- Words of caution:
  - need to fix word embeddings and composition weight matrices for the metric to have determinism/reproducibility

Distributed, Compositional Models and MT Evaluation

- Clear scope to incorporate (general-purpose) distributed word/phrase representations into MT evaluation metrics
- Also possibility to include composition into models (a la Socher et al. [2011] or Kalchbrenner and Blunsom [2013])
- Words of caution:
  - need to fix word embeddings and composition weight matrices for the metric to have determinism/reproducibility
  - slight concerns about the domain-stability of the learned representations

Continuous Syntax-based Neural SMT

- Another natural direction for distributed, compositional SMT models is neural SMT incorporating some model of syntax, relevant to which is:

Continuous Syntax-based Neural SMT

- Another natural direction for distributed, compositional SMT models is neural SMT incorporating some model of syntax, relevant to which is:
    - Socher et al. [2013]: parsing with CVGs and syntactically untied NNs (e.g. for decoding into English)

Continuous Syntax-based Neural SMT

- Another natural direction for distributed, compositional SMT models is neural SMT incorporating some model of syntax, relevant to which is:
  - Socher et al. [2013]: parsing with CVGs and syntactically untied NNs (e.g. for decoding into English)
  - Kalchbrenner and Blunsom [2013]: implicit syntactic parsing via composition with convolutional NNs

Continuous Syntax-based Neural SMT

- Another natural direction for distributed, compositional SMT models is neural SMT incorporating some model of syntax, relevant to which is:
  - Socher et al. [2013]: parsing with CVGs and syntactically untied NNs (e.g. for decoding into English)
  - Kalchbrenner and Blunsom [2013]: implicit syntactic parsing via composition with convolutional NNs
  - Jones et al. [2012]: parsing with synchronous hyperedge replacement grammars

Bringing Neural SMT to the Masses: Factored Neural SMT

- Factored SMT [Koehn and Hoang, 2007] is a famously attractive mechanism for incorporating arbitrary (linguistic) features (e.g. morphology or semantics) into an SMT system in the form of extra features in the log-linear model

Bringing Neural SMT to the Masses: Factored Neural SMT

- Factored SMT [Koehn and Hoang, 2007] is a famously attractive mechanism for incorporating arbitrary (linguistic) features (e.g. morphology or semantics) into an SMT system in the form of extra features in the log-linear model
  - despite intuitive promise and ease of use, factored SMT hard to get to work in practice

Bringing Neural SMT to the Masses: Factored Neural SMT

- Factored SMT [Koehn and Hoang, 2007] is a famously attractive mechanism for incorporating arbitrary (linguistic) features (e.g. morphology or semantics) into an SMT system in the form of extra features in the log-linear model
    - despite intuitive promise and ease of use, factored SMT hard to get to work in practice
    - neural SMT perhaps offers a more promising way of integrating arbitrary features as part of "soft" multi-order representation (cf. Socher et al. [2013], or simply as the basis for learning extra "feature embeddings")

Talk Outline

1. Elements of a Compositional, Distributed SMT Model
2. Training a Compositional, Distributed SMT Model
3. Semantics and SMT
4. Moving Forward
5. Summary

Summary

- Composed, distributed reflections on semantics and SMT, including:
    - what is a distributed representation, distributional semantics, semantic composition, and what are some standard approaches to each?
    - what bits of semantics have contributed to SMT in the past and why; what does this tell us about the recent successes of "neural SMT"?
    - random thoughts on possible short- to medium-term possibilities for research on semantic SMT

Summary

- Composed, distributed reflections on semantics and SMT, including:
  - what is a distributed representation, distributional semantics, semantic composition, and what are some standard approaches to each?
  - what bits of semantics have contributed to SMT in the past and why; what does this tell us about the recent successes of "neural SMT"?
  - random thoughts on possible short- to medium-term possibilities for research on semantic SMT
  - what is semantics anyway?

Acknowledgements

- Thanks to the University of Melbourne NLP Group for their patience and valuable input on an earlier, very sketchy version of this talk
- This work was supported by the Australian Research Council

References I

Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 597–604, Ann Arbor, USA, 2005.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 238–247, Baltimore, USA, 2014. URL http://www.aclweb.org/anthology/P14-1023.

Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, pages 50–57, Toulouse, France, 2001. URL http://www.aclweb.org/anthology/P01-1008.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 11–21, Atlanta, USA, 2013. URL http://www.aclweb.org/anthology/S13-1002.

Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261, 2005.

References II

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Ondřej Bojar and Jan Hajič. Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, USA, 2008. URL `http://www.aclweb.org/anthology/W/W08/W08-0319`.

Francis Bond. *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*. CSLI Publications, Stanford, USA, 2005.

Francis Bond, Stephan Oepen, Eric Nichols, Dan Flickinger, Erik Velldal, and Petter Haugereid. Deep open-source machine translation. *Machine Translation*, 25(2): 87–105, 2011.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th International Global Wordnet Conference (GWC 2012)*, Matsue, Japan, 2012.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

Clara Cabezas and Philip Resnik. Using WSD techniques for lexical selection in statistical machine translation. Technical Report CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42, University of Maryland, 2005.

References III

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Czech Republic, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1007.

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 (ACL 2013)*, pages 1435–1445, Sofia, Bulgaria, 2013. URL http://www.aclweb.org/anthology/P13-1141.

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 for Computational Linguistics (ACL 2007)*, pages 33–40, Prague, Czech Republic, 2007. URL http://www.aclweb.org/anthology/P07-1005.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Ido Dagan and Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596, 1994.

References IV

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. TESLA at WMT 2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 78–84, Edinburgh, UK, 2011. URL http://www.aclweb.org/anthology/W11-2106.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, UK, 2011. URL http://www.aclweb.org/anthology/W11-2107.

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 (ACL 2014)*, pages 1370–1380, Baltimore, USA, 2014. URL http://www.aclweb.org/anthology/P14-1129.

Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the ACL and 3rd Annual Meeting of the NAACL (ACL-02)*, pages 255–262, Pittsburgh, USA, 2002.

Georgiana Dinu and Mirella Lapata. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1162–1172, Cambridge, USA, 2010. URL http://www.aclweb.org/anthology/D10-1113.

References V

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 350–356, Geneva, Switzerland, 2004.

Bonnie J. Dorr. The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7:135–193, 1992/3.

Bonnie J. Dorr. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322, 1997.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 115–119, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/P12-2023.

Edward Grefenstette. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 1–10, Atlanta, USA, 2013. URL http://www.aclweb.org/anthology/S13-1001.

G.E. Hinton, J.L. McClelland, and D.E. Rumelhart. Distributed representations. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 77–109. MIT Press, Cambridge, USA, 1986.

References VI

Eric Huang, Richard Socher, Christopher Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/P12-1092.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. *Nihongo Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten, 1997.

Bevan Jones, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. Semantics-based machine translation with hyperedge replacement grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1359–1376, Mumbai, India, 2012. URL http://www.aclweb.org/anthology/C12-1083.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1700–1709, Seattle, USA, 2013. URL http://www.aclweb.org/anthology/D13-1176.

Kevin Knight and Steve K. Luk. Building a large-scale knowledge base for machine translation. In *Proceedings of the 12th Annual Conference on Artificial Intelligence (AAAI-94)*, pages 773–778, Seattle, USA, 1994. URL arxiv.org/abs/cmp-lg/9407029.

References VII

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*, pages 868–876, Prague, Czech Republic, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1091.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France, 2012.

Hai-Son Le, Alexandre Allauzen, and François Yvon. Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2012)*, pages 39–48, Montréal, Canada, 2012. URL http://www.aclweb.org/anthology/N12-1005.

Mike Lewis and Mark Steedman. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192, 2013. URL http://aclweb.org/anthology//Q/Q13/Q13-1015.pdf.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 354–359, Uppsala, Sweden, 2010. URL http://www.aclweb.org/anthology/W10-1754.

References VIII

Chi-kiu Lo and Dekai Wu. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 220–229, Portland, USA, 2011. URL http://www.aclweb.org/anthology/P11-1023.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252, Montréal, Canada, 2012. URL http://www.aclweb.org/anthology/W12-3129.

Marco Lui, Timothy Baldwin, and Diana McCarthy. Unsupervised estimation of word usage similarity. In *Proceedings of the Australasian Language Technology Workshop 2012 (ALTW 2012)*, pages 33–41, Dunedin, New Zealand, 2012.

Thang Luong, Richard Socher, and Christopher D. Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the 17th Conference on Natural Language Learning (CoNLL-2013)*, pages 104–113, Sofia, Bulgaria, 2013. URL http://www.aclweb.org/anthology/W13-3512.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Japan, 2010.

References IX

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA, 2013.

Teruko Mitamura, Eric H Nyberg, and Jaime G Carbonell. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of the Third Machine Translation Summit (MT Summit III)*, 1991.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010.

Hiromi Nakaiwa, Satoshi Shirai, Satoru Ikehara, and T. Kawaoka. Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints. In *Proceedings of the AAAI Spring Symposium Series, Empirical Methods in Discourse Interpretation and Generation*, pages 99–105, 1995.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 455–462, Sapporo, Japan, 2003. URL `http://www.aclweb.org/anthology/P03-1058`.

Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

References X

Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, USA, 1993. URL http://www.aclweb.org/anthology/P93-1024.

Lee Schwartz, Takako Aikawa, and Chris Quirk. Disambiguation of English PP attachment using multilingual aligned data. In *Proceedings of the Ninth Machine Translation Summit (MT Summit IX)*, New Orleans, USA, 2003.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809, 2011.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1201–1211, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/D12-1110.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 455–465, Sofia, Bulgaria, 2013. URL http://www.aclweb.org/anthology/P13-1045.

References XI

Dan Tufiș, Radu Ion, and Nancy Ide. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1312–1318, Geneva, Switzerland, 2004.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. Word-sense disambiguation for machine translation. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 771–778, Vancouver, Canada, 2005. URL http://www.aclweb.org/anthology/H/H05/H05-1097.

Julie Weeds, David Weir, and Diana McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1015–1021, Geneva, Switzerland, 2004.

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, and Shouxun Lin. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 750–758, Jeju Island, Korea, 2012. URL http://www.aclweb.org/anthology/P12-1079.

Deyi Xiong and Min Zhang. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1459–1469, Baltimore, USA, 2014. URL http://www.aclweb.org/anthology/P14-1137.

References XII

Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems (NIPS 2007)*, pages 1689–1696, Vancouver, Canada, 2007.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1393–1398, Seattle, USA, 2013. URL `http://www.aclweb.org/anthology/D13-1141`.