Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable

Viktor Hangya[1], Fabienne Braune[1,2], Alexander Fraser[1], Hinrich Schütze[1]

[1]Center for Information and Language Processing
LMU Munich, Germany
[2]Volkswagen Data Lab Munich, Germany
{hangyav, fraser}@cis.uni-muenchen.de
fabienne.braune@volkswagen.de

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 640550).

Introduction

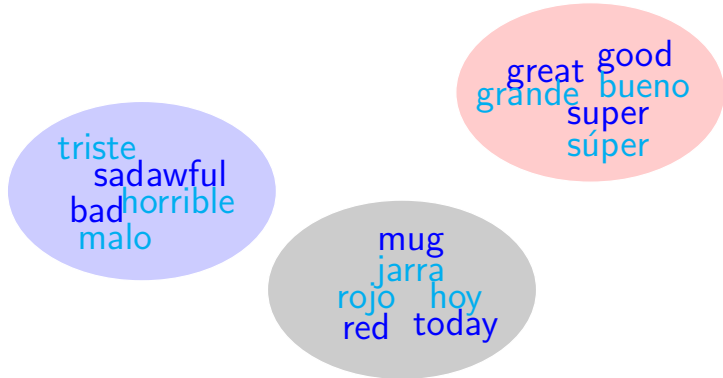
- Bilingual transfer learning is important for overcoming data sparsity in the target language
- Bilingual word embeddings eliminate the gap between source and target language vocabulary
- Resources required for bilingual methods are often out-of-domain:
  - Texts for embeddings
  - Source language training samples
- We focused on domain-adaptation of word embeddings and better use of unlabeled data

Motivation

- Cross-lingual sentiment analysis of tweets

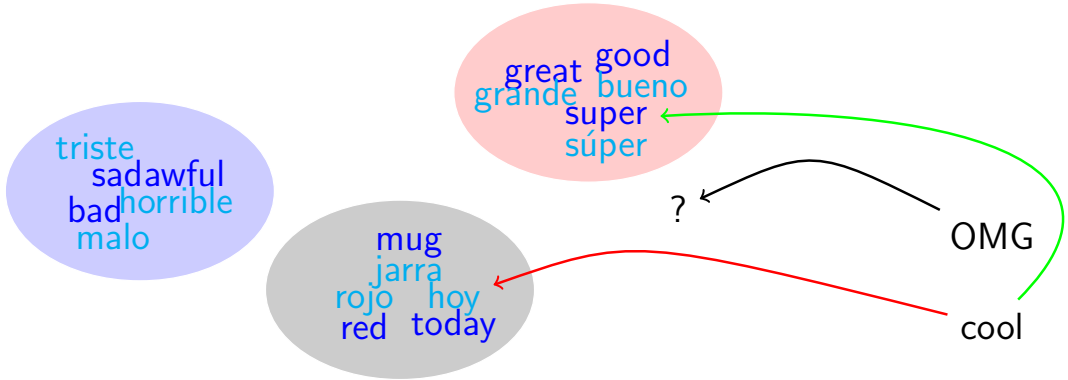
Motivation

- Cross-lingual sentiment analysis of tweets



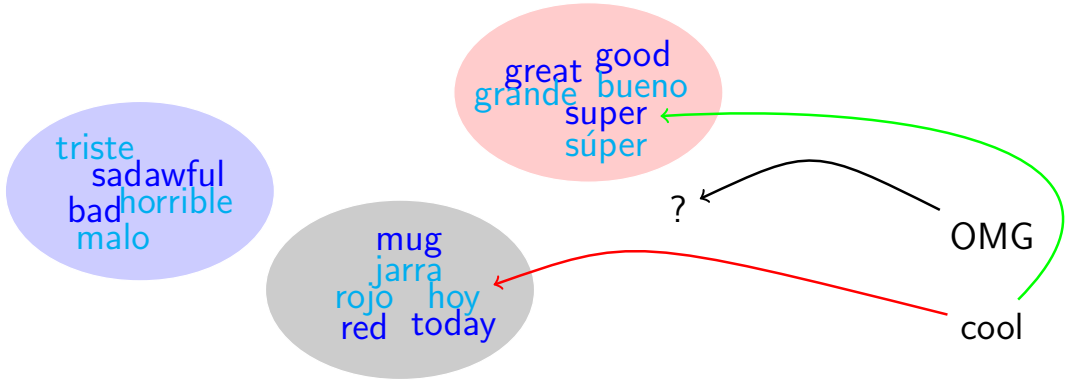
Motivation

- Cross-lingual sentiment analysis of tweets



Motivation

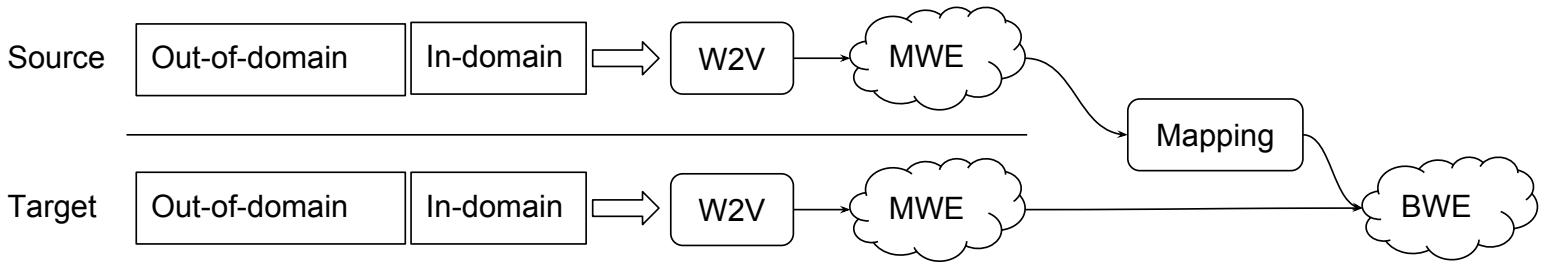
- Cross-lingual sentiment analysis of tweets



- Combination of two methods:
  - Domain adaptation of bilingual word embeddings
  - Semi-supervised system for exploiting unlabeled data

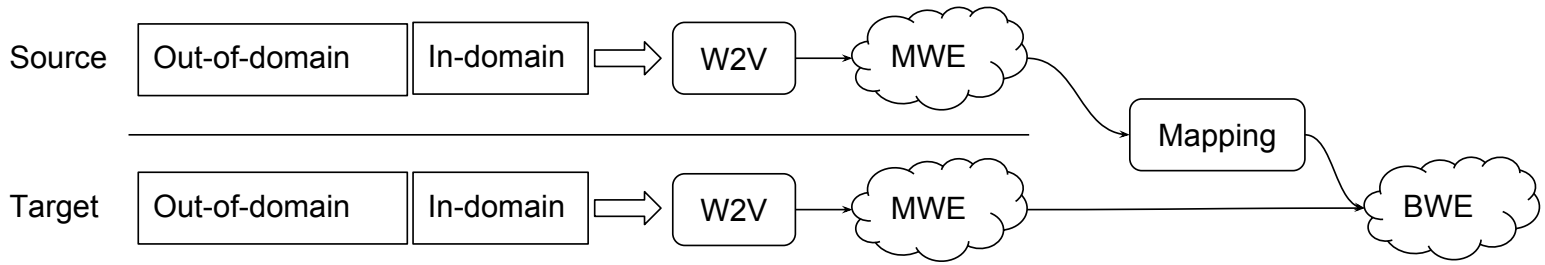
- No additional annotated resource is needed:
  - Cross-lingual sentiment classification of tweets
  - Medical bilingual lexicon induction

Word Embedding Adaptation



- Goal: domain-specific bilingual word embeddings with general domain semantic knowledge

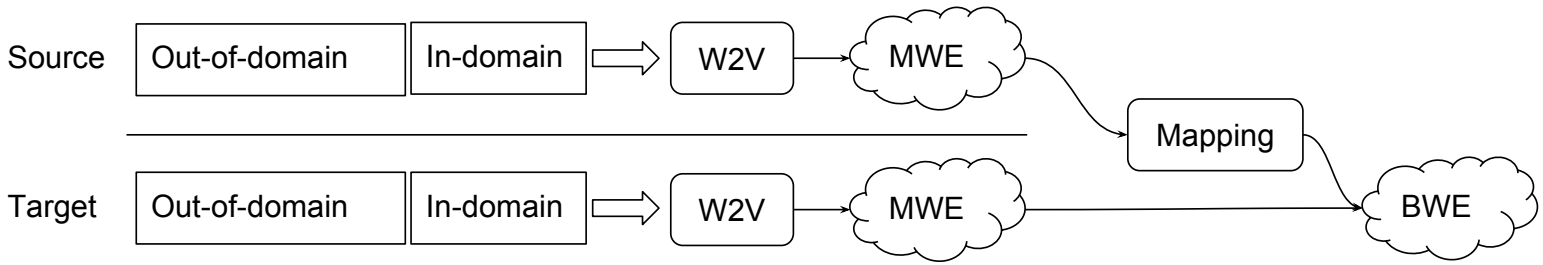
Word Embedding Adaptation



- Goal: domain-specific bilingual word embeddings with general domain semantic knowledge

1. Monolingual word embeddings on concatenated data (Mikolov et al., 2013):
   - Easily accessible general (out-of-domain) data
   - Domain-specific data

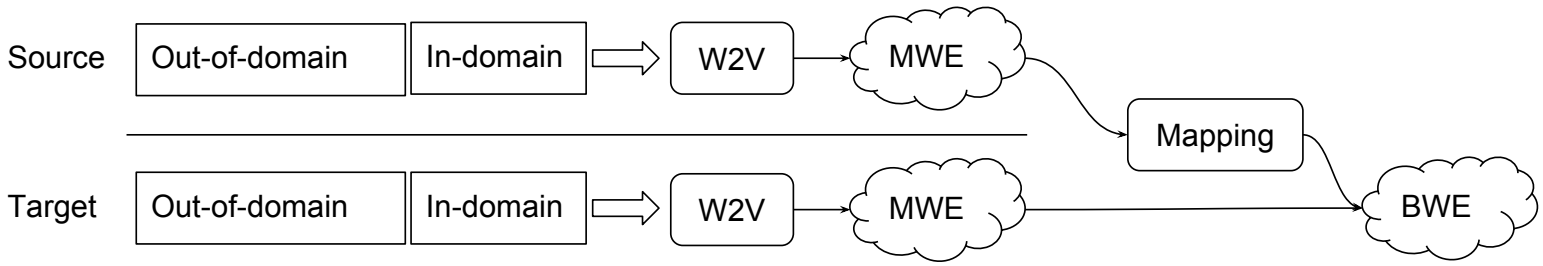
Word Embedding Adaptation



- Goal: domain-specific bilingual word embeddings with general domain semantic knowledge

1. Monolingual word embeddings on concatenated data (Mikolov et al., 2013):
   - Easily accessible general (out-of-domain) data
   - Domain-specific data
2. Map monolingual embeddings to a common space using post-hoc mapping (Mikolov et al., 2013)
   - Small seed lexicon containing word pairs is needed

Word Embedding Adaptation



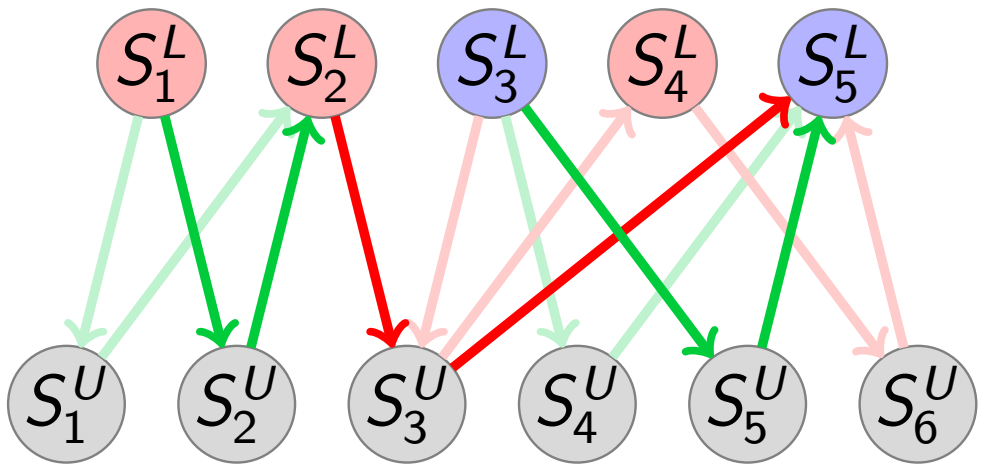
- Goal: domain-specific bilingual word embeddings with general domain semantic knowledge

1. Monolingual word embeddings on concatenated data (Mikolov et al., 2013):
   - Easily accessible general (out-of-domain) data
   - Domain-specific data
2. Map monolingual embeddings to a common space using post-hoc mapping (Mikolov et al., 2013)
   - Small seed lexicon containing word pairs is needed

- **Simple and intuitive but crucial for the next step!**

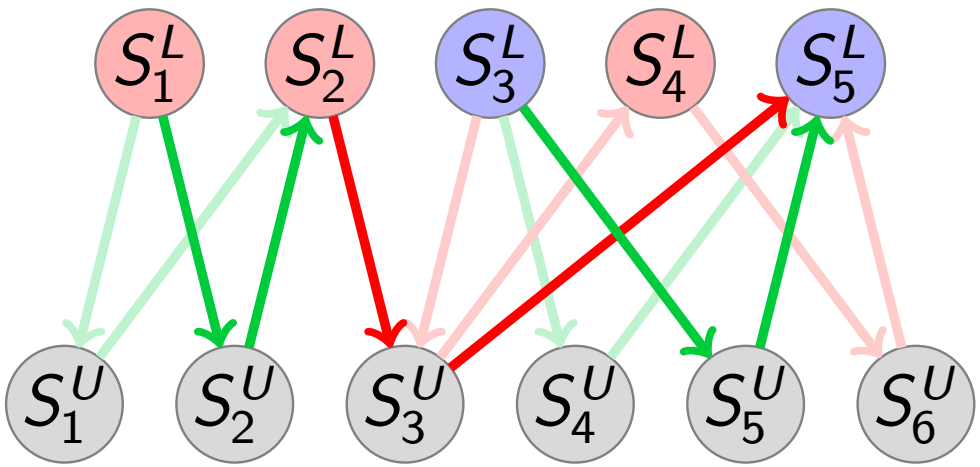
Semi-Supervised Approach

- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \lambda_2 * \mathcal{L}_{walker} + \lambda_3 * \mathcal{L}_{visit}$



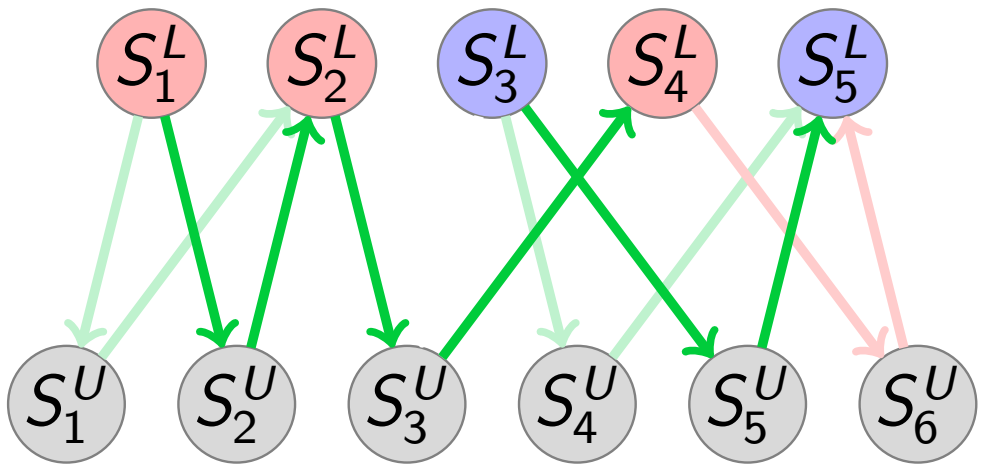
Semi-Supervised Approach

- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \boldsymbol{\lambda_2 * \mathcal{L}_{walker}} + \lambda_3 * \mathcal{L}_{visit}$



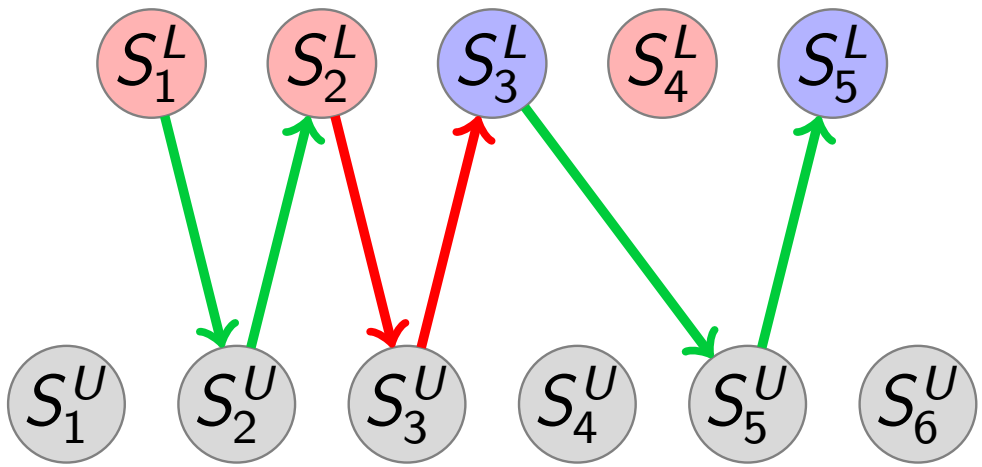
Semi-Supervised Approach

- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \boldsymbol{\lambda_2 * \mathcal{L}_{walker}} + \lambda_3 * \mathcal{L}_{visit}$



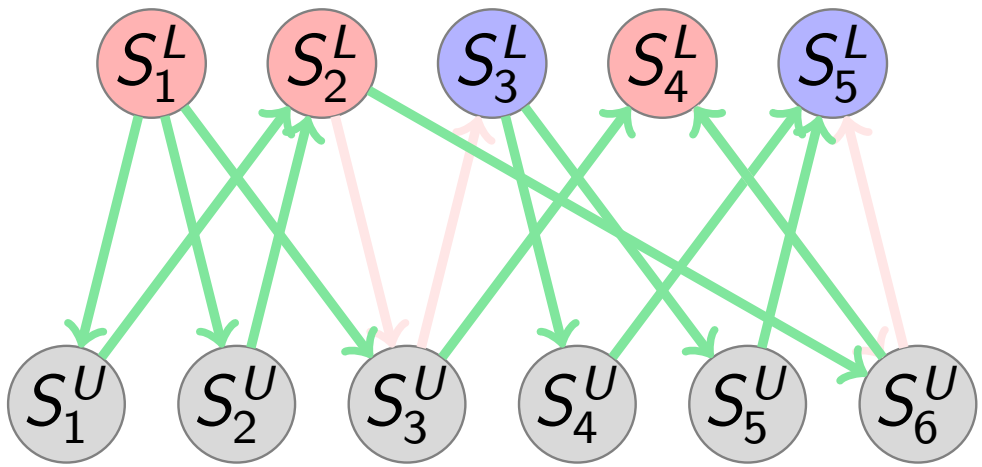
Semi-Supervised Approach

- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \lambda_2 * \mathcal{L}_{walker} + \boldsymbol{\lambda_3 * \mathcal{L}_{visit}}$



Semi-Supervised Approach

- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \lambda_2 * \mathcal{L}_{walker} + \boldsymbol{\lambda_3 * \mathcal{L}_{visit}}$



Semi-Supervised Approach

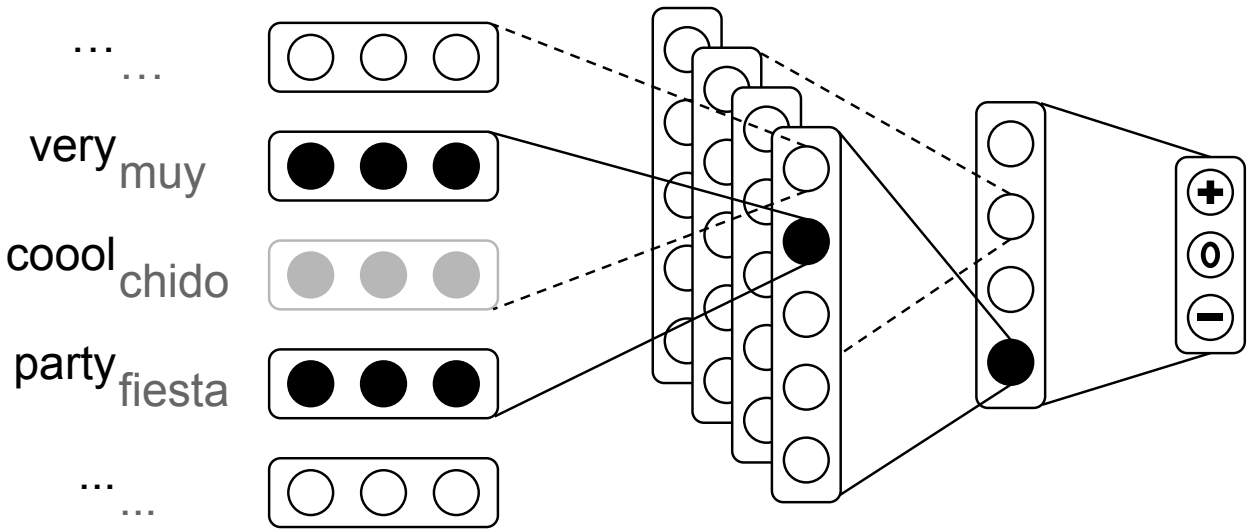
- Goal: Unlabeled samples for training
- Tailored system from computer vision to NLP (Häusser et al., 2017)
  - Labeled/unlabeled samples in the same class are similar
  - Sample representation is given by the $n-1^{th}$ layer
  - Walking cycles: labeled $\rightarrow$ unlabeled $\rightarrow$ labeled
  - Maximize the number of correct cycles
  - $\mathcal{L} = \lambda_1 * \mathcal{L}_{classification} + \lambda_2 * \mathcal{L}_{walker} + \lambda_3 * \mathcal{L}_{visit}$
- Adapted bilingual word embeddings make the models able to find correct cycles at the beginning of the training and improve them later on.

Cross-Lingual Sentiment Analysis of Tweets

- RepLab 2013 sentiment classification (+/0/-) of En/Es tweets (Amigó et al., 2013)
  - @churcaballero jajaja con lo bien que iba el volvo...
- General domain data: 49.2M OpenSubtitles sentences (Lison and Tiedemann, 2016)
- Twitter specific data:
  - 22M downloaded tweets
  - RepLab Background
- Seed lexicon: frequent English words from BNC (Kilgarriff, 1997)
- Labeled data: RepLab En training set
- Unlabeled data: RepLab Es training set

Cross-Lingual Sentiment Analysis of Tweets

- Our method is easily applicable to word embedding-based off-the-shelf classifiers



CNN classifier

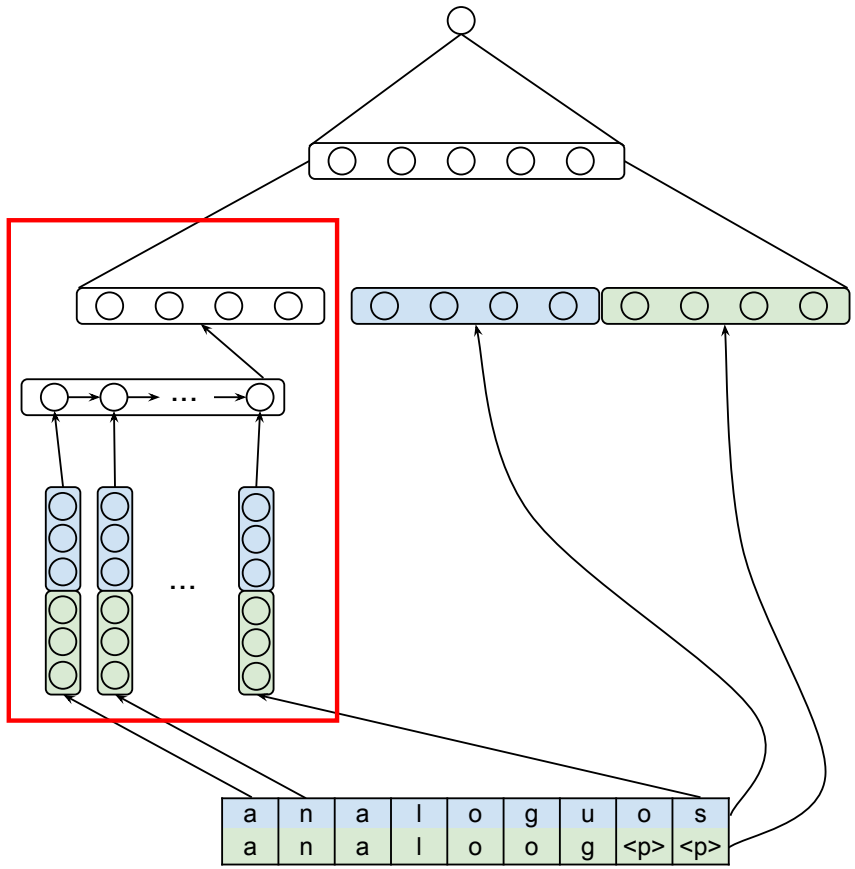
(Kim, 2014)

Medical Bilingual Lexicon Induction

- Mine Dutch translations of English medical words
  (Heyman et al., 2017)
  - *sciatica* $\rightarrow$ *ischias*
- General domain data: 2M Europarl (v7) sentences
- Medical data: 73.7K medical Wikipedia sentences
- Medical seed lexicon (Heyman et al., 2017)
- Unlabeled
  1. En word in BNC $\rightarrow$ 5 most similar and 5 random Du pair
  2. En word in medical lexicon $\rightarrow$ 3 most similar Du $\rightarrow$
     $\rightarrow$ 5 most similar and 5 random En

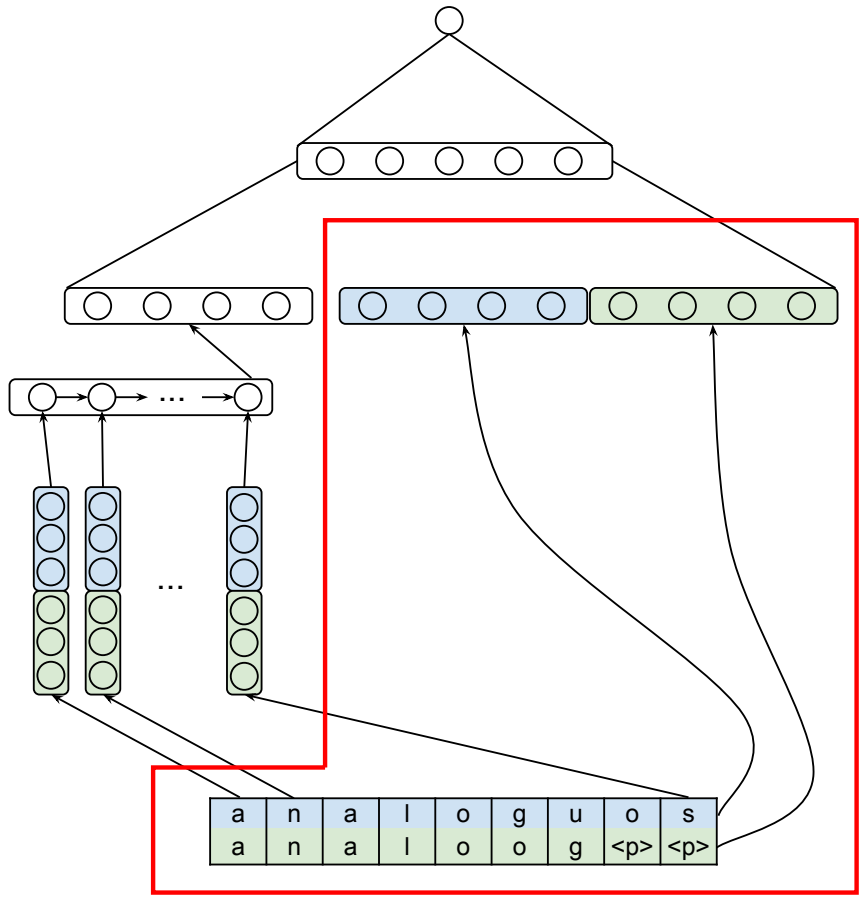
Medical Bilingual Lexicon Induction

- Classifier based approach (Heyman et al., 2017)
  - Word pairs as training set (negative sampling)
  - Character level LSTM to learn orthographic similarity



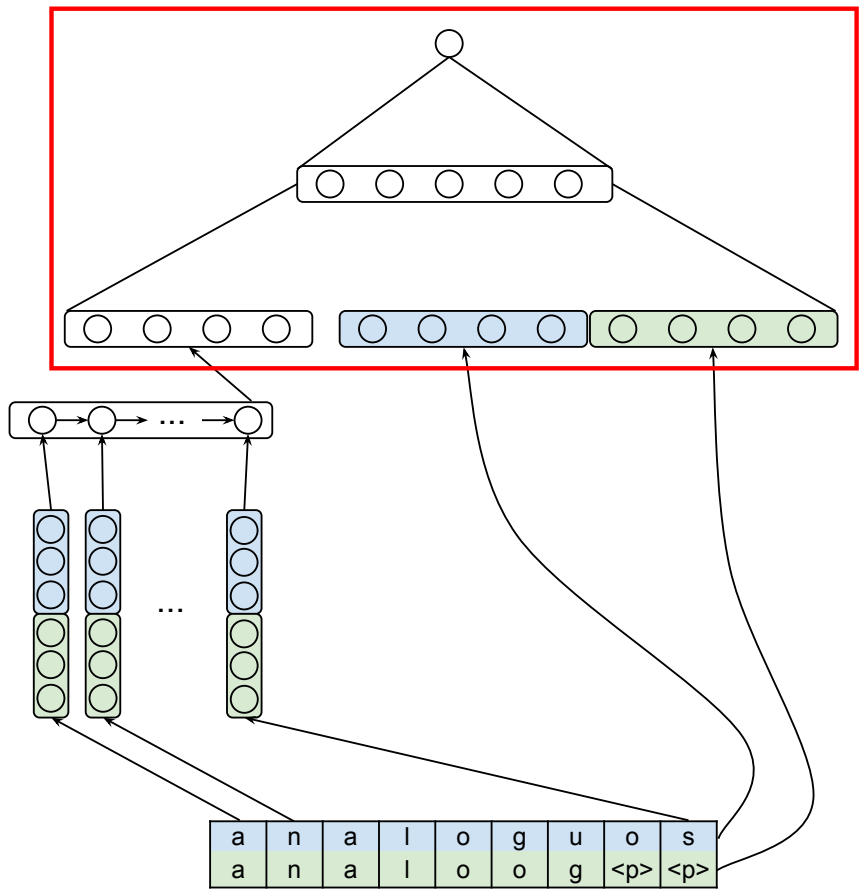
Medical Bilingual Lexicon Induction

- Classifier based approach (Heyman et al., 2017)
  - Word pairs as training set (negative sampling)
  - Word embeddings to learn semantic similarity



Medical Bilingual Lexicon Induction

- Classifier based approach (Heyman et al., 2017)
  - Word pairs as training set (negative sampling)
  - Dense-layer scores word pairs



Results: Sentiment Analysis

labeled data	En
unlabeled data	-
Baseline	59.05%
BACKGROUND	58.50%
22M_tweets	**61.14%**
Subtitle+BACKGROUND	59.34%
Subtitle+22M_tweets	**61.06%**

Table 1: Accuracy on cross-lingual sentiment analysis of tweets

Results: Sentiment Analysis

labeled data unlabeled data	En -	En Es
Baseline	59.05%	58.67% (-0.38%)
BACKGROUND	58.50%	57.41% (-1.09%)
22M_tweets	61.14%	60.19% (-0.95%)
Subtitle+BACKGROUND	59.34%	**60.31% (0.97%)**
Subtitle+22M_tweets	61.06%	**63.23% (2.17%)**

Table 1: Accuracy on cross-lingual sentiment analysis of tweets

Results: Sentiment Analysis

labeled data	En	En	En+Es
unlabeled data	-	Es	-
Baseline	59.05%	58.67% (-0.38%)	-
BACKGROUND	58.50%	57.41% (-1.09%)	-
22M_tweets	61.14%	60.19% (-0.95%)	-
Subtitle+BACKGROUND	59.34%	60.31% **(0.97%)**	62.92% **(2.61%)**
Subtitle+22M_tweets	61.06%	63.23% **(2.17%)**	63.82% **(0.59%)**

Table 1: Accuracy on cross-lingual sentiment analysis of tweets

Results: Bilingual Lexicon Induction

labeled lexicon	medical	BNC
unlabeled lexicon	-	-
Baseline	35.70	20.73
Europarl+Medical	**40.71**	**22.10**

Table 2: F_1 scores of medical bilingual lexicon induction

Results: Bilingual Lexicon Induction

labeled lexicon	medical	BNC	medical	medical
unlabeled lexicon	-	-	medical	BNC
Baseline	35.70	20.73	**36.20 (0.50)**	35.04 (-0.66)
Europarl+Medical	40.71	22.10	**41.44 (0.73)**	**41.01 (0.30)**

Table 2: F_1 scores of medical bilingual lexicon induction

Conclusions

- Bilingual transfer learning yield poor results when using out-of-domain resource
- We showed that performance can be increased by using only additional unlabeled monolingual data
  - Delightfully simple approach to adapt embeddings
  - Broadly applicable method to exploit unlabeled data
- Language and task independent approaches

Thank your for your attention!

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 640550).

References

[1] Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, Damiano Spina, Enrique Amigo, Jorge Carrillo de Albornoz, Tamara Martin, and Maarten de Rijke. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Proc. CLEF*.

[2] Philip Häusser, Alexander Mordvintsev, and Daniel Cremers. 2017. Learning by Association - A versatile semi-supervised training method for neural networks. In *Proc. CVPR*.

[3] Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proc. EACL*.

[4] Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*.

[5] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. EMNLP*.

[6] Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proc. LREC*.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. ICLR*.

[8] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.