Inducing Crosslingual Distributed Representations of Words

Alexandre Klementiev Ivan Titov Binod Bhattarai
{aklement, titov, bhattara}@mmci.uni-saarland.de

Motivation

With large vocabularies and limited annotated data treating words as atomic symbols often means poor model estimates.

Instead, we can induce alternative representations from cheap unsupervised data and use them instead of words:

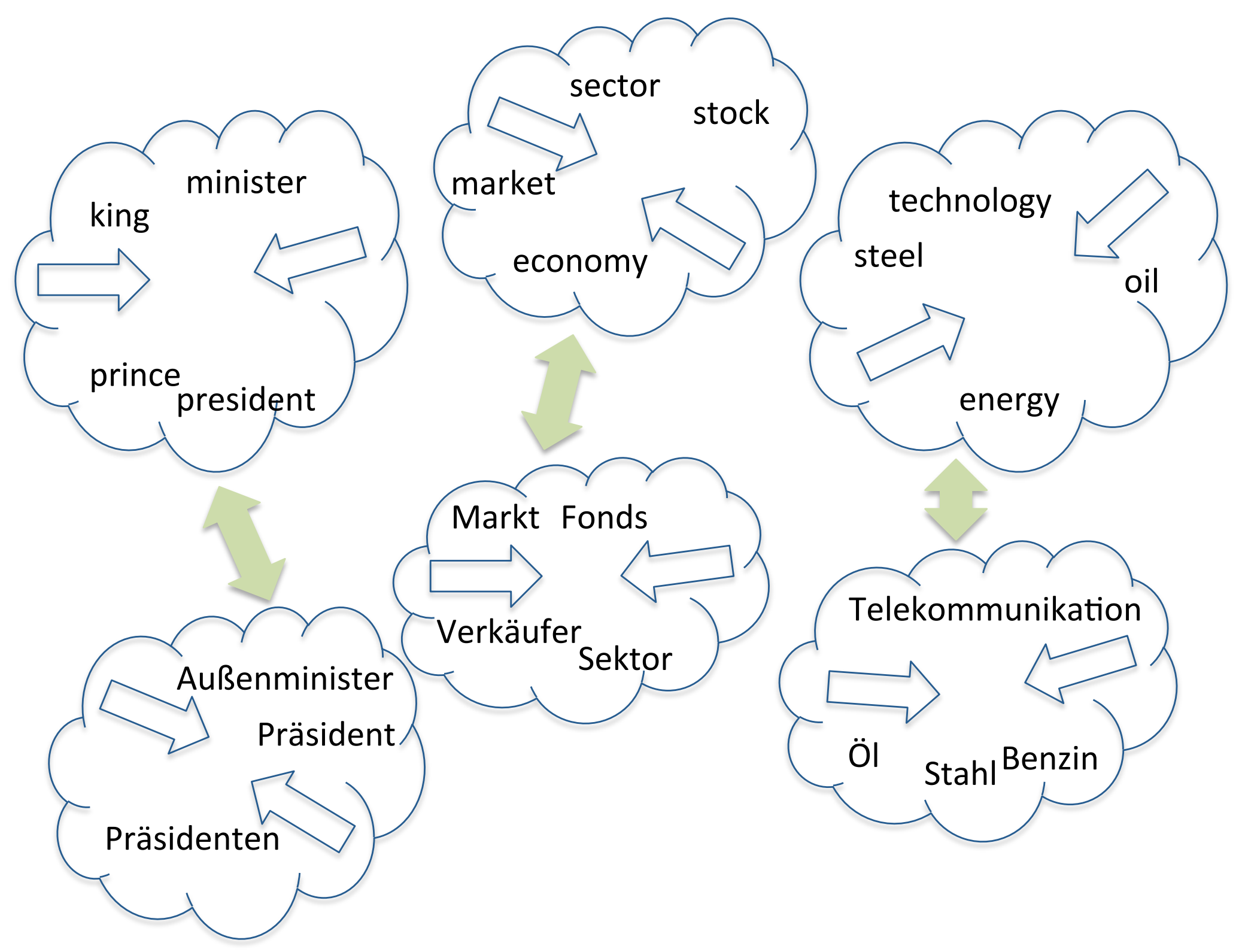
Clustering	Vector space	Distributed
○ Assign words to (hierarch.) clusters ○ Words defined by cluster prototypes	○ Words defined by context	○ Vector space + probabilistic models ○ Dense embedding
How to choose granularity?	Algorithmically induced	Low dimensional
Many clusterings possible	Not learned for a given task	Learned (for a given task)

Inducing the same representation for a pair of languages has additional benefits for low resource languages. We can learn in one language where annotation is available and apply to the other language directly.

Our Contribution

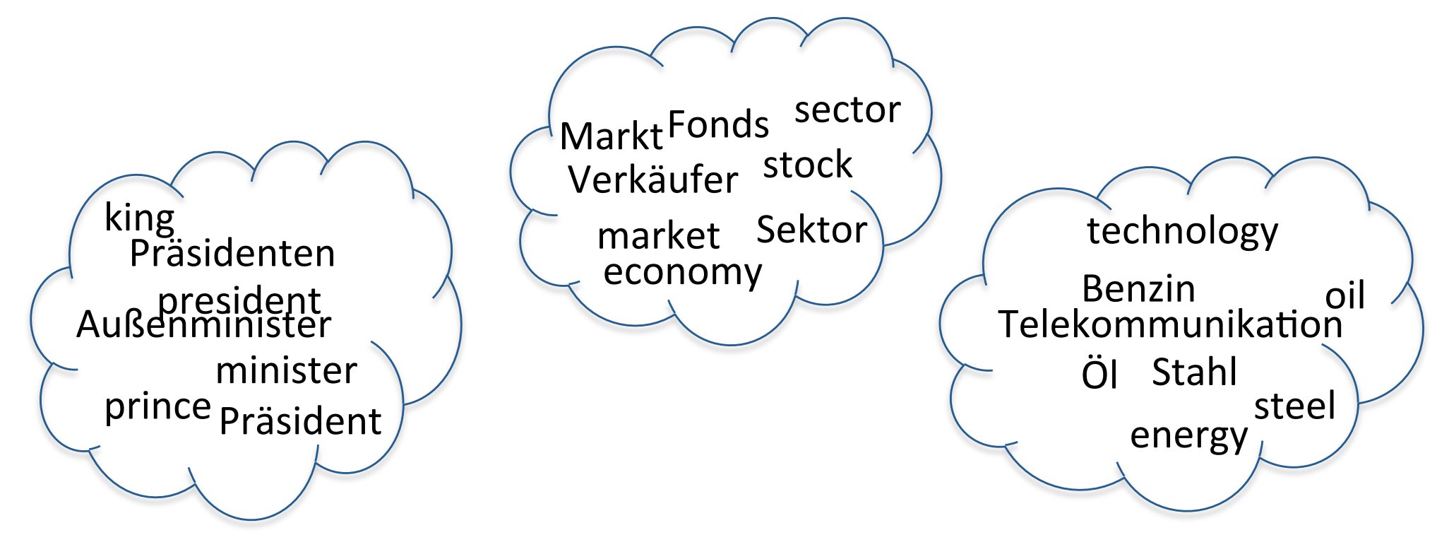
A general multitask learning (MTL) inspired framework to induce crosslingual distributed representations.

- Treat words as individual tasks
- Task relatedness is derived from co-occurrence statistics in bilingual parallel data



Learn joint representation using:

- ⇨ monolingual data to induce a representation within each language
- ⇔ parallel data to bias representations to be similar for translated words



Multitask Learning

The goal of Multitask Learning (MTL) is to improve generalization performance across a set of tasks by learning them jointly

The MTL setup of Cavallanti et al. (2010):

- Consider K tasks
- Learn a linear classifier parameterized by c_k for each task (c is the concatenated parameter vector)
- Minimize the following objective:

$$L(c) = \sum_{k=1}^{K} L^{(k)}(c_k) + \frac{1}{2} c^\top (A \otimes I_m) c$$

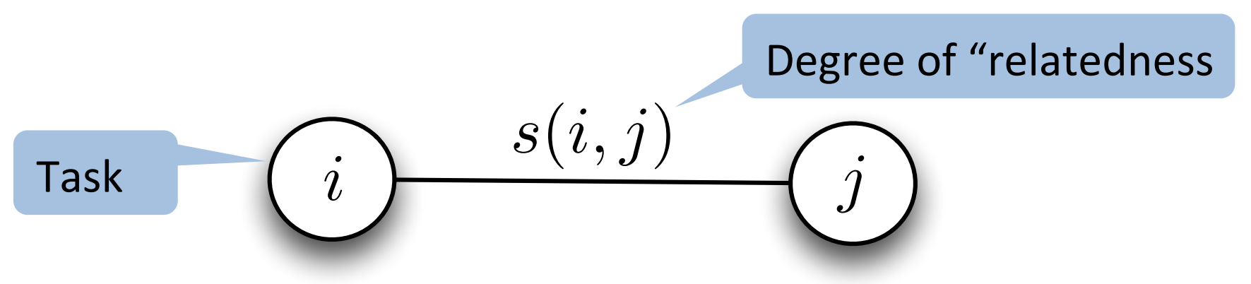
Loss function for task k

Regularizer prefers "similar" parameters for related tasks

- Interaction matrix A encodes task "relatedness"

Encoding prior knowledge into the interaction matrix:

- Represent tasks with an undirected weighted graph $H = (R, E, S)$:



- The graph Laplacian is defined as:

$$J_{i,j}(H) = \begin{cases} \sum_{(i,k)\in E} s(i,k) & \text{if } i = j \\ -s(i,j) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

- Interaction matrix is then defined as $A = I + J$
- A^{-1} defines the degree of "relatedness" between tasks
- A is invertible (J is positive semi-definite)

Crosslingual Representation Induction

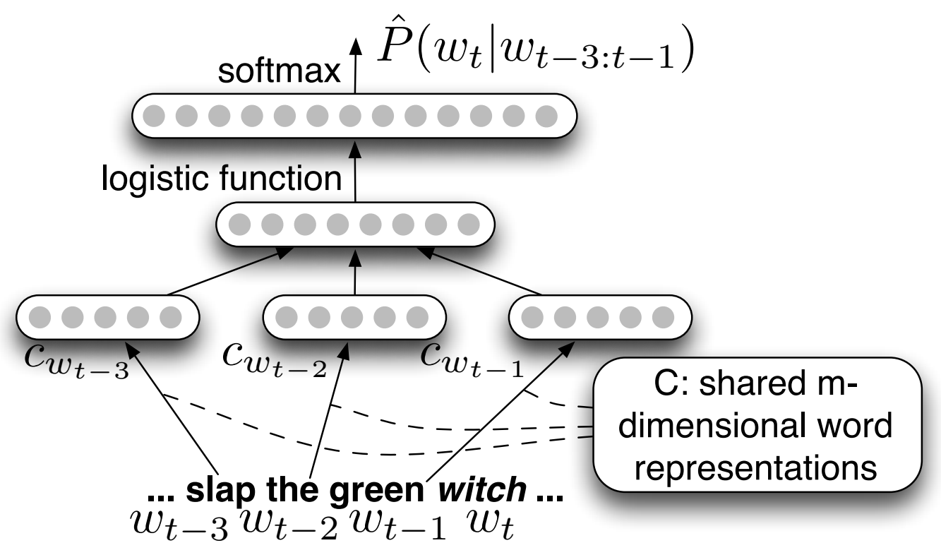
Idea: frame crosslingual distributed representation induction as multi-task learning

- Treat words in both languages as individual tasks
- Define A by how often words align in parallel data
- Use the regularizer from the MTL objective

Applicable to any distributed representation induction set-up

- In this work, we apply it to the neural probabilistic language model [Bengio et al. (2003)]

- Re-formulate the learning objective:


softmax
$\hat{P}(w_t|w_{t-3:t-1})$
logistic function
$c_{w_{t-3}}$
$c_{w_{t-2}}$
$c_{w_{t-1}}$
C: shared m-dimensional word representations
... slap the green witch ...
w_{t-3} w_{t-2} w_{t-1} w_t

Includes c

$$L(\theta) = \sum_{l=1}^{2} \sum_{t=1}^{T^{(l)}} \log \hat{P}_{\theta^{(l)}}(w_t^{(l)} | w_{t-n+1:t-1}^{(l)}) + \frac{1}{2} c^\top (A \otimes I_m) c$$

Over both languages

Language modeling part captures intra-language word similarities

Regularizer ensures crosslingual similarity in the induced embedding

- Train using stochastic gradient descent
- Computing A^{-1} is hard, use an approximation

Evaluation

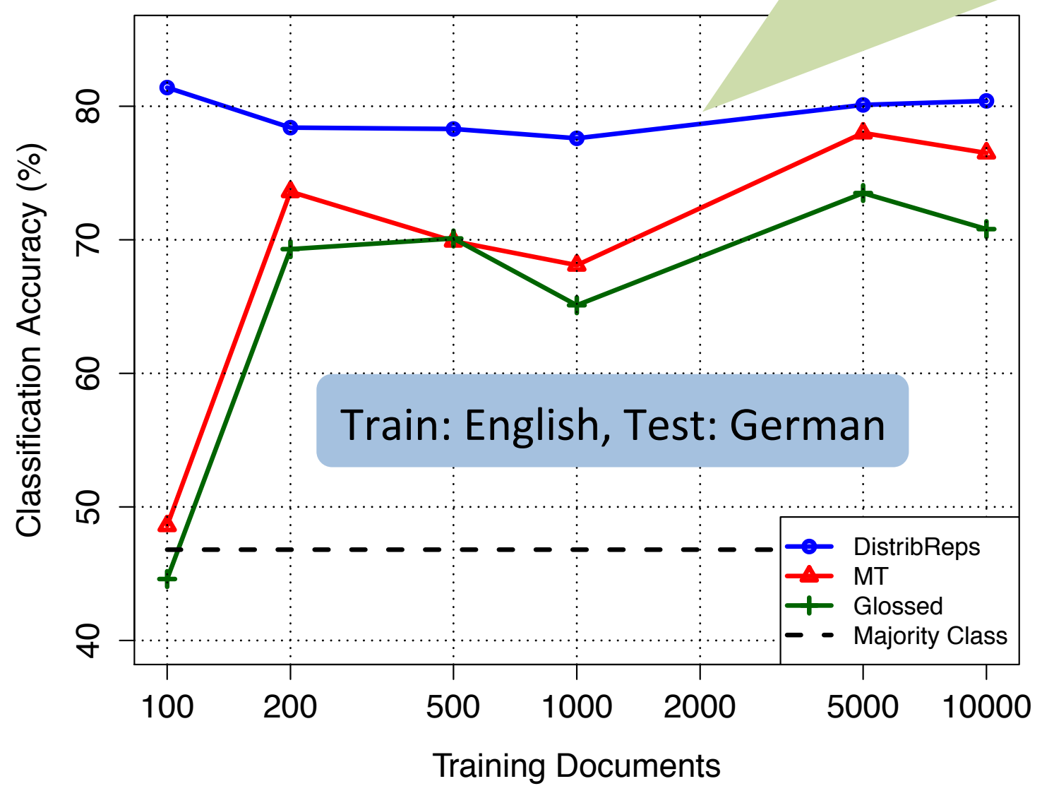
- Induced 40-dimensional representation of words in German and English
- Used RCV1/2 monolingual corpora (~8 million tokens in each language)
- Used Europarl parallel data to define the interaction matrix

Crosslingual document classification (4 RCV document topics):

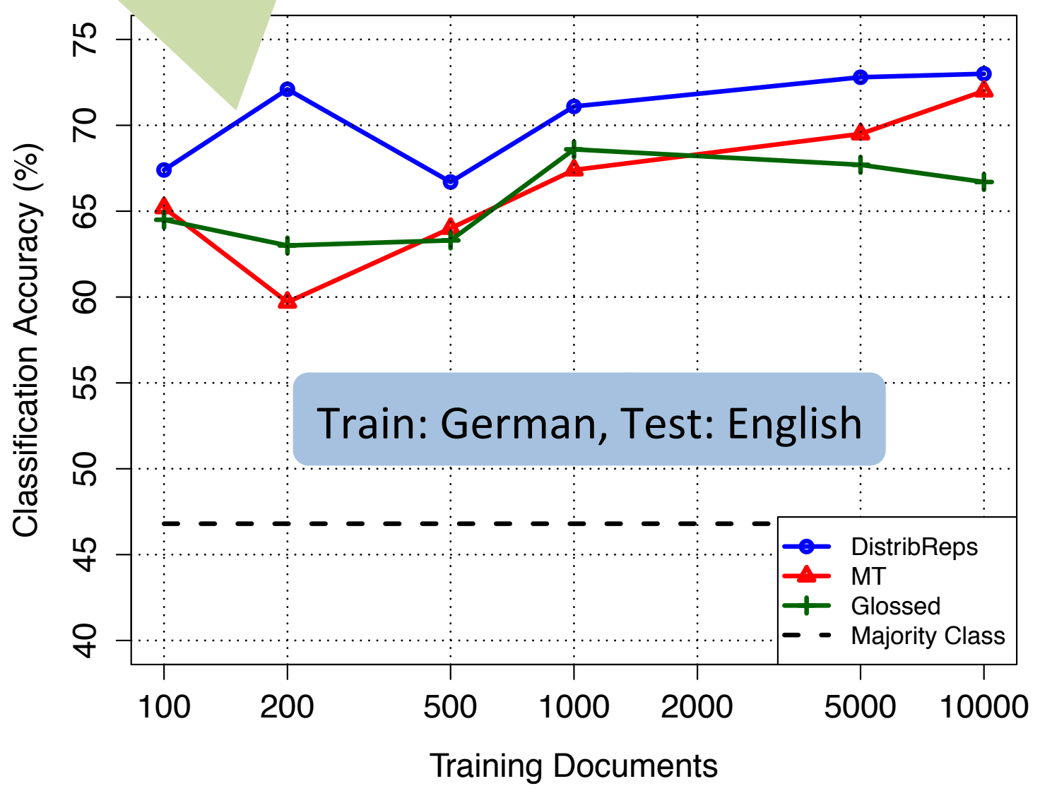
- Train on annotated data in one language
- DistribReps: test on second language directly with no additional training
- Glossed: translate test docs into the original language (most frequently aligned words)
- MT: same but with phrase-based MT
- Results are likely to improve with increased embedding dimensionality

january		president		said	
january	januar	president	präsident	said	sagte
february	februar	king	präsidenten	reported	erklärte
november	november	hun	minister	stated	sagten
april	april	areas	staatspräsident	told	meldete
august	august	saddam	hun	declared	berichtete
march	märz	minister	vorsitzenden	stressed	sagt
june	juni	advisers	us-präsident	informed	ergänzte
december	dezember	prince	könig	announced	erklärten
july	juli	representative	berichteten	explained	teilt
september	september	institutional	außenminister	warned	berichteten

Crosslingual representations are very informative



Train: English, Test: German



Train: German, Test: English

- Embedding can be learned for a specific task by incorporating a discriminative term in the objective

The work was supported by the MMCI Cluster of Excellence and a Google research award.