Inducing Crosslingual Distributed Representations of Words

Alex Klementiev, Ivan Titov and Binod Bhattarai

CLUSTER OF EXCELLENCE

UNIVERSITÄT DES SAARLANDES

Motivation: Word Representations

- NLP systems treating words as atomic symbols need a lot of annotated data:
  - I.e. vectors with a single one, and many zeros
  - But vocabs are large, many words are rare

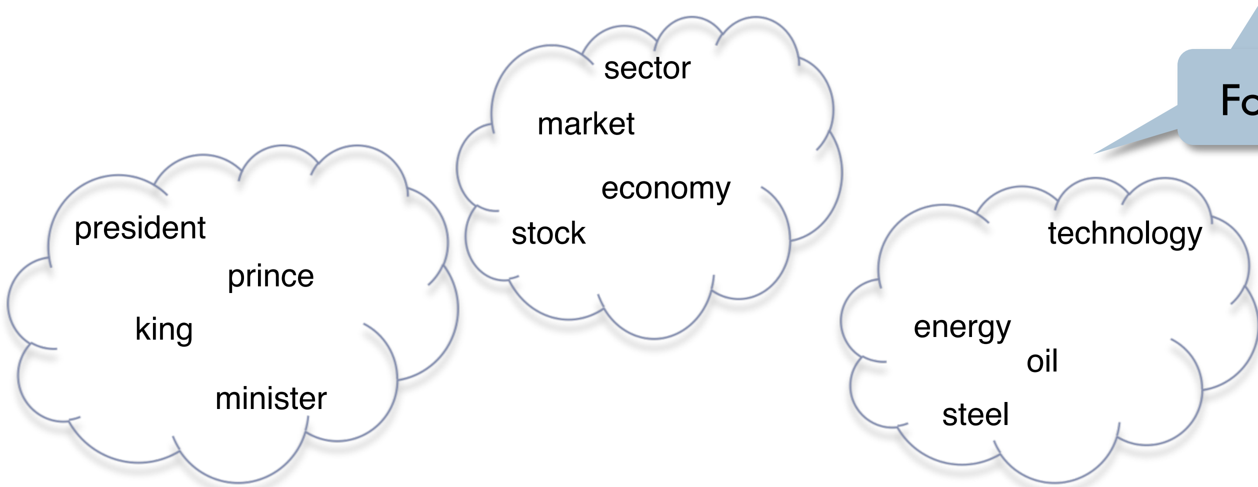
Poor model estimates

- Can address this by inducing representations for words instead
  - Use cheap unsupervised data to induce them
  - Use them as features for a learning task
- Very effective on a number of NLP tasks
  - Dependency parsing [Koo et.al., 2008], NER [Turian et.al., 2010],…

Motivation: Distributed Representations

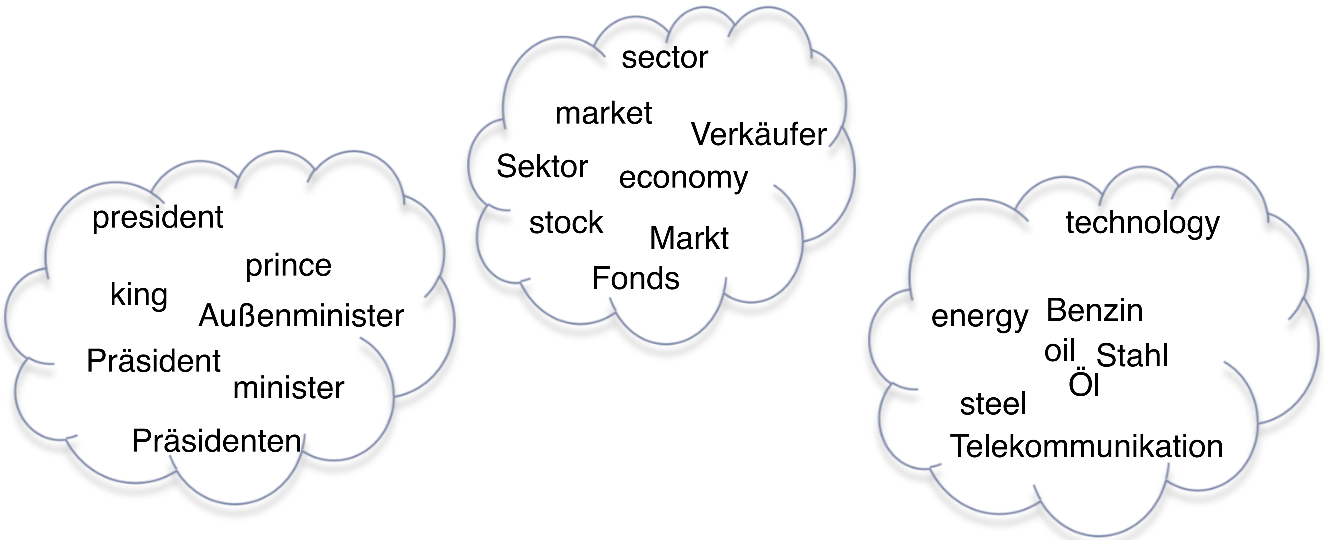
Clustering	Vector space	Distributed
‣ Cluster words into (hierarchical) clusters ‣ Words defined by cluster prototypes	‣ Words defined by context	‣ Vector space + probabilistic models ‣ Dense embedding
How to choose granularity?	Algorithmically induced	Low dimensional
Many clusterings possible		Learned (for a given task)

Focus of this work



Why Crosslingual Representations?

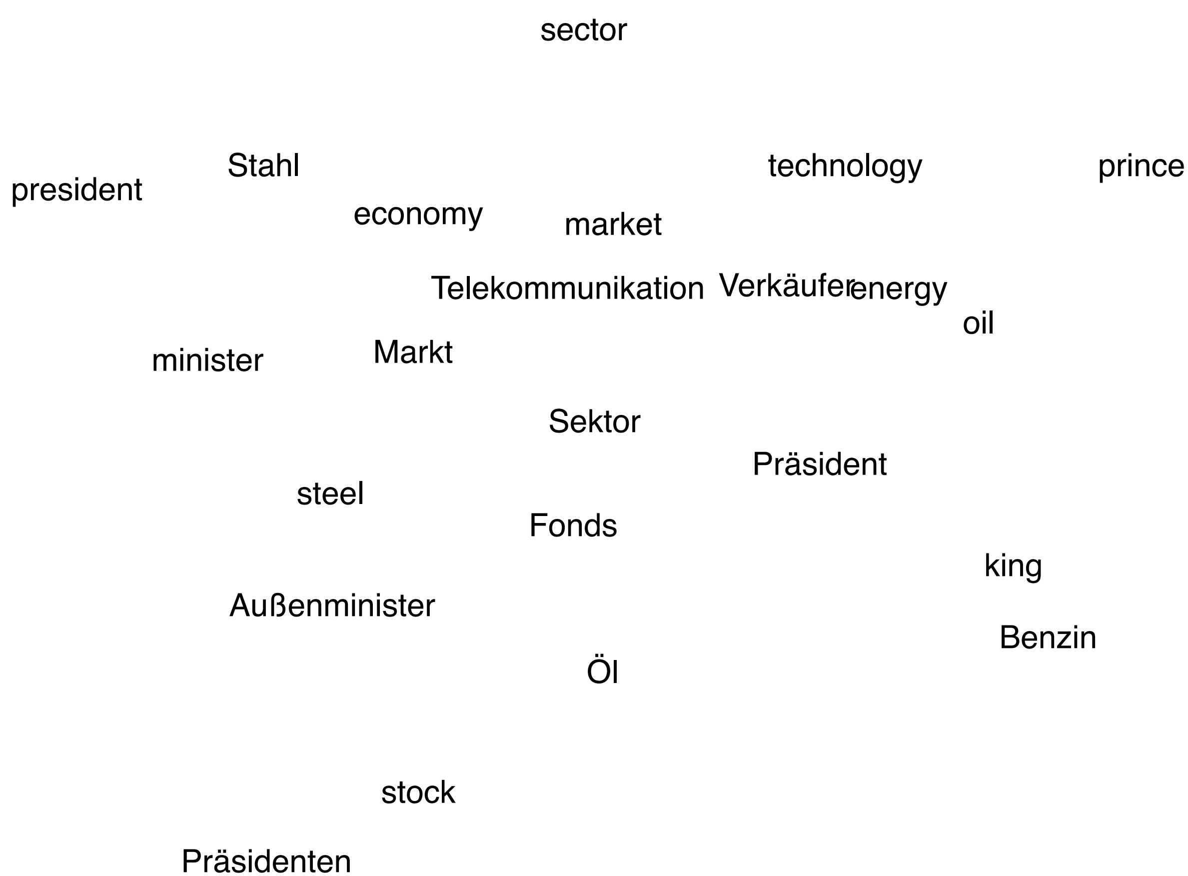
- *Same* representation for both languages:



- Especially important when one of the languages is low resource
  - Learn in one language where annotation is available – apply to the other *directly*!

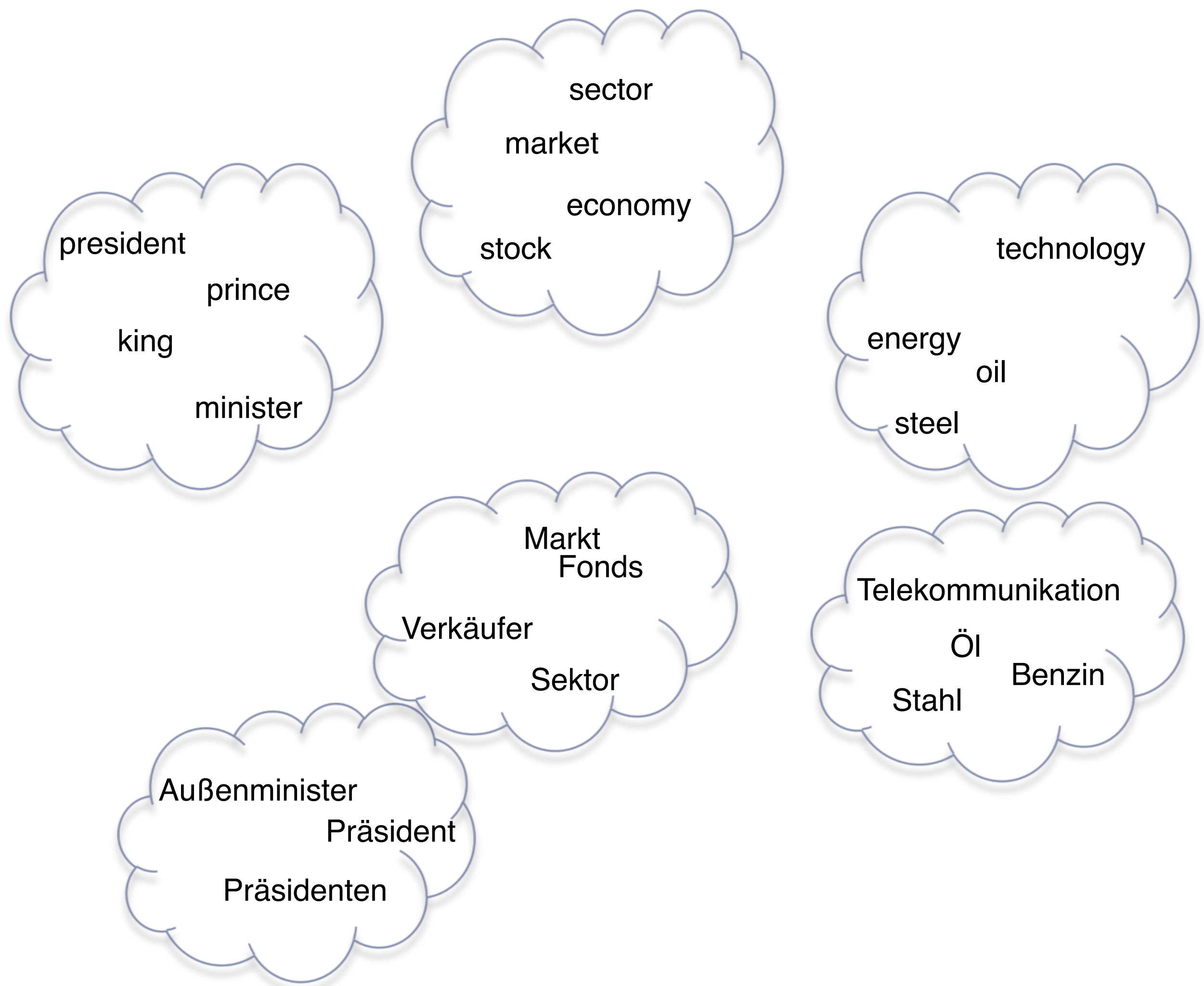
Our contribution: a general multitask learning inspired framework to induce crosslingual distributed representations

Summary of our Approach



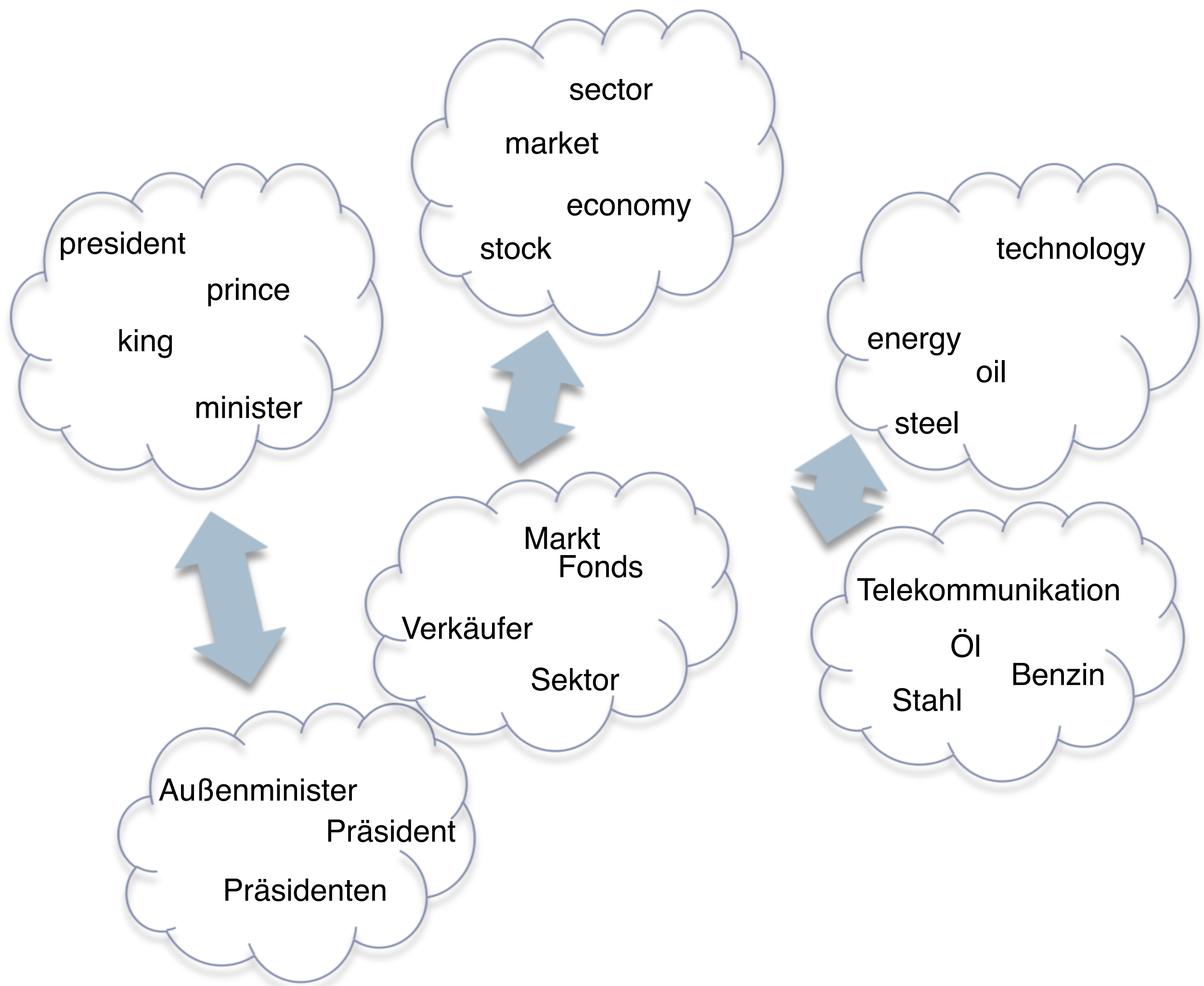
Summary of our Approach

- Use cheap monolingual data to induce a representation within each language



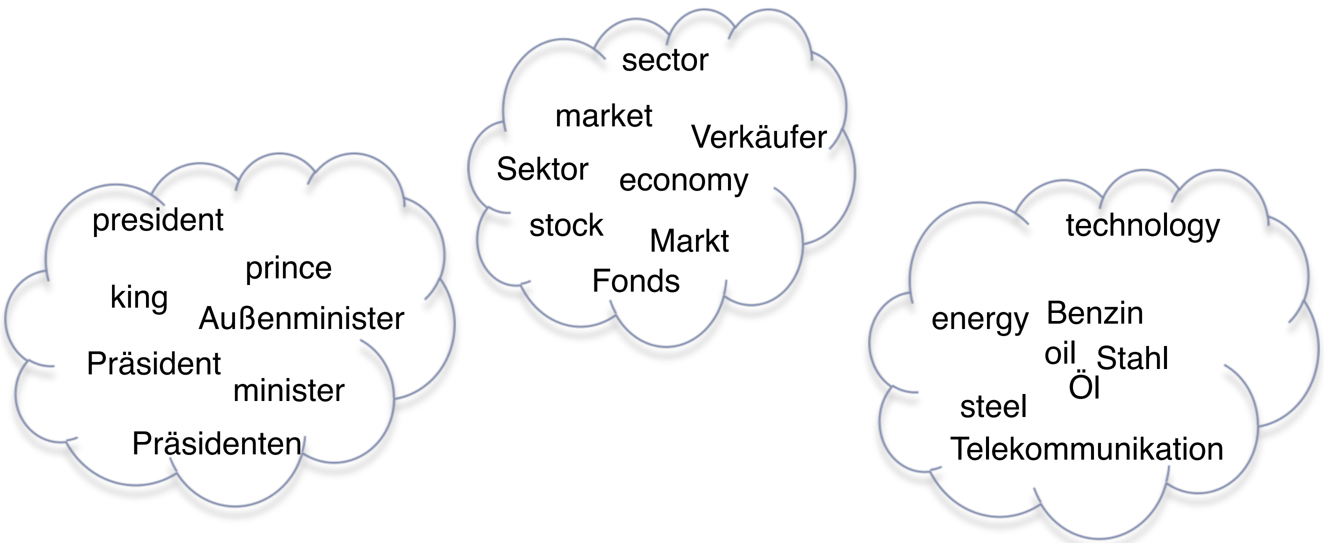
Summary of our Approach

- While using parallel data to bias representations to be similar for translated words



Summary of our Approach

- Semantically similar words are "close" to one another irrespective of language



- Treat it as multitask learning (MTL)
  - Treat words as individual tasks
  - Task relatedness is derived from co-occurrence statistics in bilingual parallel data

This work is first to address crosslingual distributed representation induction

Outline

- Motivation and summary of the approach
- Background
  - Multitask learning
  - Neural Language Models
- Crosslingual Distributed Representation Induction
- Experiments
  - Qualitative Evaluation
  - Applications to Crosslingual Document Classification

Background: Multitask Learning

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

- Idea: learn related tasks together using a shared representation
- Intuition: information is propagated across tasks
- Particularly useful when sufficient annotation is not available for (some of) the tasks

Background: Multitask Learning

- We consider a particular MTL setup [Cavallanti et al. (2010)]
- Consider K tasks; a multitask learner receives a labeled example at time *t* for one of the tasks:

$x_t \in \mathbb{R}^m$ — Example

y_t — Correct Label

$i_t \in [1, K]$ — Task index

- Learns a linear classifier (parameterized by $v_j, j \in [1, K]$) for each task
- Minimizes the following objective:

Defines inter task similarity

$$L(v) = \sum_t L^{(t)}(v_{i_t}) + R(v, A)$$

Prefers "similar" parameters for related tasks

Background: Multitask Learning

- For multitask binary perceptron, the objective corresponds to:

$$v_j \leftarrow v_j + y_t A^{-1}_{j,i_t} x_t$$

Rate of update for tasks related to i_t

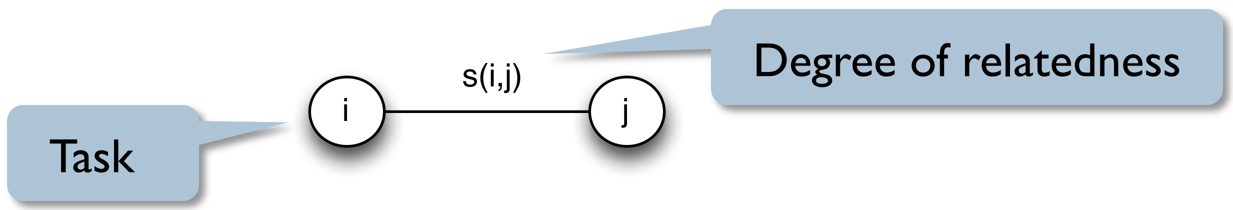
- When a mistake is made, updates are distributes to all related tasks
- Interaction matrix *A* defines task "relatedness", e.g.:

$$A^{-1} = \frac{1}{K+1} \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix}$$

All tasks are equally related to other tasks

Background: Multitask Learning

- How can we encode prior knowledge of task relatedness into A?
- Represent tasks with an undirected weighted graph *H*:



- The graph *Laplacian L* is defined as:

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

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

What do we take from MLT?

Our idea: frame crosslingual distributed representation induction as multi-task learning

- We treat words in both languages as individual tasks
- We will take the the multitask regularizer part of the objective

$$L(v) = \sum_t L^{(t)}(v_{i_t}) + \boxed{R(v, A)}$$

$$\boxed{\frac{1}{2} v^\top (A \otimes I_m) v}$$

- Applicable to any distributed representation induction set-up

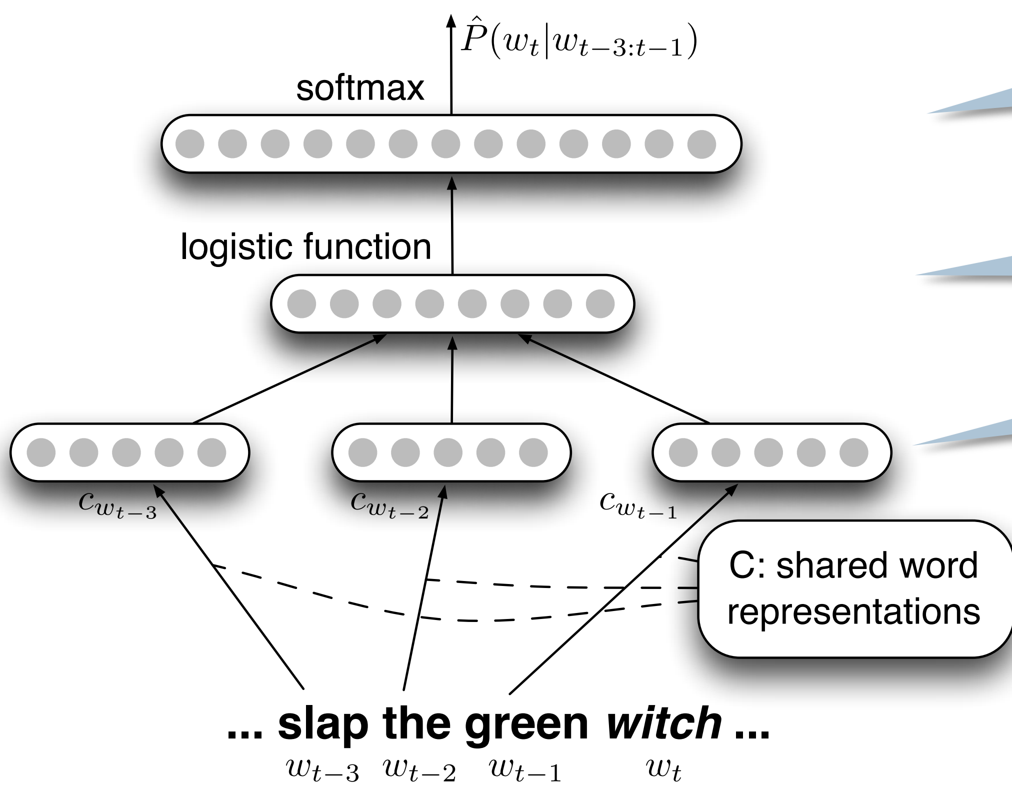
In this work, we apply it to neural language models (next)

Outline

- Motivation and summary of the approach
- Background
  - Multitask learning
  - Neural Language Models
- Crosslingual Distributed Representation Induction
- Experiments
  - Qualitative Evaluation
  - Applications to Crosslingual Document Classification

Background: Neural Distributed Representations

Neural probabilistic models learn a latent multi-dimensional representation of words and use them to estimate the probability distribution of word sequences



Turn into prob. distribution (a node for each word)

Apply linear transformation followed by logistic function

Concatenate representations

Map context words to shared representation

Key component!

Background: Neural Distributed Representations

- An important side-effect of training NLMs are the d-dimensional *shared representation c*:
  - Capture semantic properties of context words, because these properties are predictive of a possible next word
  - Induced vectors are "closer" for more similar words
  - Learned with other parameters using backpropagation
- Learning maximizes the following objective:

$$L(\theta) = \boxed{\sum_{t=1}^{T} \log \hat{P}_\theta(w_t | w_{t-n+1:t-1})}$$

c and other parameters

Outline

- Motivation and summary of the approach
- Background
  - Multitask learning
  - Neural Language Models
- Crosslingual Distributed Representation Induction
- Experiments
  - Qualitative Evaluation
  - Applications to Crosslingual Document Classification

Crosslingual Representation Induction

Goal: Induce an embedding so that semantically similar words are "close" irrespective of the language

- Train neural language models *jointly* to induce a *common* embedding
  - Use monolingual data in each language to induce representations
- Use the MTL framework to ensure crosslingual similarity
  - Use parallel data to define the interaction matrix A

Crosslingual Representation Induction

- We formulate the learning objective as:

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

Over both languages

Language modeling part

MTL regularizer part

- Language modeling part captures intra-language word similarities
- Regularizer part ensures crosslingual similarity in the induced embedding c
- Train using stochastic gradient descent
- Representations of context words (in each language) and of words related to them are modified at each step

Defining the interaction matrix *A*

- The interaction matrix *A* defines relatedness between tasks (words)
- Use parallel data:
  - A set of sentences and their translations
  - Alignments induced with standard MT tools (GIZA++)
  - More alignments between a pair of words – more "related" they are
- Can define *A* using graph Laplacian of the (bi-partite) graph
  - Nodes are words, edge weights – number of alignments
  - However, computing inverse is expensive, use a heuristic to define A^{-1} directly:

$$\hat{A}^{-1}_{w,w'} = \frac{s(w,w')}{m_w + 1 + \sum_{\tilde{w}} s(w,\tilde{w})} \qquad \hat{A}^{-1}_{w,w} = \frac{m_w + 1}{m_w + 1 + \sum_{\tilde{w}} s(w,\tilde{w})}$$

Outline

- Motivation and summary of the approach
- Background
  - Multitask learning
  - Neural Language Models
- Crosslingual Distributed Representation Induction
- Experiments
  - Qualitative Evaluation
  - Applications to Crosslingual Document Classification

Evaluation

- Data/Setup
  - Induce 40-dimensional representation of words in German and English
  - RCV1/2 monolingual corpora (~8 million tokens in each language)
  - Europarl parallel data to define the interaction matrix
- Qualitative evaluation
  - Look at a handful of words and their closest neighbors in both languages
- Evaluation on crosslingual document classification
  - Show that the induced representations are informative
  - Evaluated on 4 class classification

Qualitative Evaluation

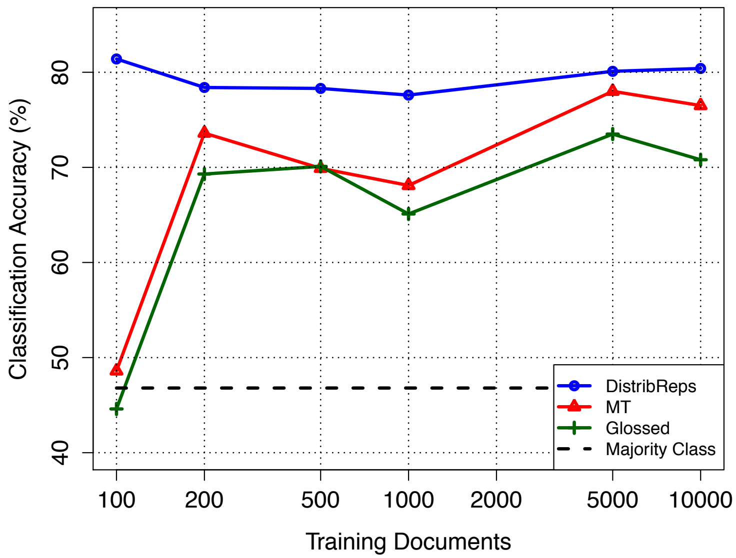
january		*president*		*said*	
en	**de**	**en**	**de**	**en**	**de**
january	januar	president	präsident	said	sagte
february	februar	king	präsidenten	reported	erklärte
november	november	hun	minister	stated	sagten
april	april	areas	staatspräsident	told	meldete
august	august	saddam	hun	declared	berichtete
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

oil		*microsoft*		*market*	
en	**de**	**en**	**de**	**en**	**de**
oil	baumwolle	microsoft	microsoft	market	markt
car	kaffee	intel	intel	papers	marktes
energy	telekommunikation	instrument	chemikalien	side	fonds
air	tabak	chapman	endesa	economy	sektor
tobacco	rindfleisch	endesa	kabel	duration	laufzeit
steel	öl	distillates	hewlett-packard	sector	montreal
housing	benzin	pty	guinness	tobacco	verkäufer
cotton	stahl	hewlett-packard	dienste	montreal	papiere
insurance	strom	guinness	thomson	house	fracht
technology	milch	potash	exxon	pay	hersteller

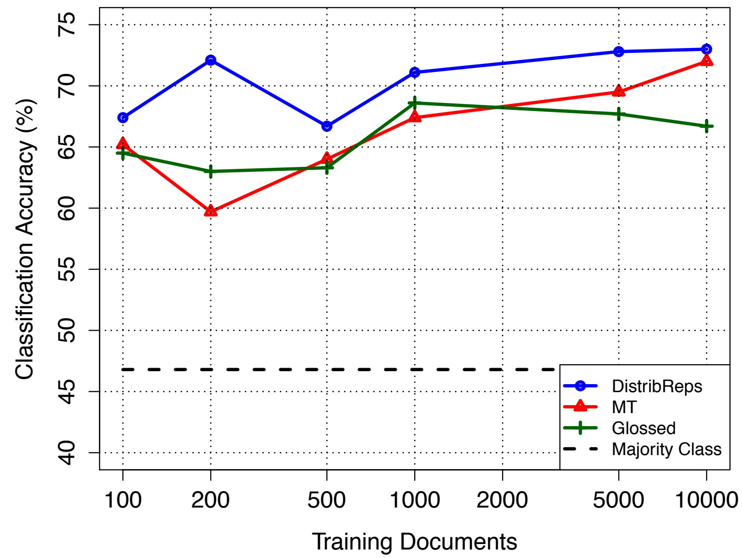
Crosslingual Document Classification

- Use distributed representations to train a classifier in one language (L1)
- Apply to the other language (L2) with *no* additional training (*DistribReps*)
- Baselines:
  - Train in L1, gloss test documents from L2 to L1 (*Glossed*)
  - Train in L1, translate (phrase-based MT) test documents in L2 to L1 (*MT*)

No training data in L2!!!



Train: en, Test: de



Train: de, Test: en

Summary and Future Work

- Proposed a general MTL-inspired framework to induce crosslingual distributed representations
  - Use cheap monolingual data to induce representation
  - Use parallel data to define a regularizer to "align" two languages
- Show that representations are very informative
  - Crosslingual document classification
- Future work
  - How sensitive the representations are to the amount of parallel data?
  - Representations of phrases: useful for low resource MT, etc.