Multilingual Models for Compositional Distributed Semantics

Karl Moritz Hermann and Phil Blunsom

Presented by: Darren Foong

parallel corpora

Multilingual Models for Compositional Distributed Semantics

parallel corpora

representations of sentences, documents etc.

Multilingual Models for **Compositional** Distributed Semantics

parallel corpora

representations of sentences,
documents etc.

Multilingual Models for Compositional Distributed Semantics

word embeddings,
vectors etc.

“... given enough parallel data, a shared representation of two parallel sentences would be forced to capture the common elements between these two sentences.”

The Idea

- Generate word embeddings (not just English) such that:
  - representations of semantically equivalent **sentences** are similar
  - representations of semantically different **sentences** are dissimilar
  - ...in parallel corpora
- Can extend to documents

The Approach

- Given functions $f : X \to \mathbb{R}^d, g : Y \to \mathbb{R}^d$
  - map **sentences** in language X and Y to representations

  and parallel corpus $C \subseteq X \times Y$
- Define "energy" of model for $(x, y) \in C$
  - $E_{bi}(x, y) = \|f(x) - g(y)\|^2$
  - Idea: minimise energy for all $(x, y) \in C$

The Approach

- Add noise-contrastive large-margin update
  - ensures representations of non-aligned sentences observe a certain margin from each other
- For each $(x, y) \in C$ sample $(x, n) \in C$
  - where x, n are not semantically equivalent (with high probability)

The Approach

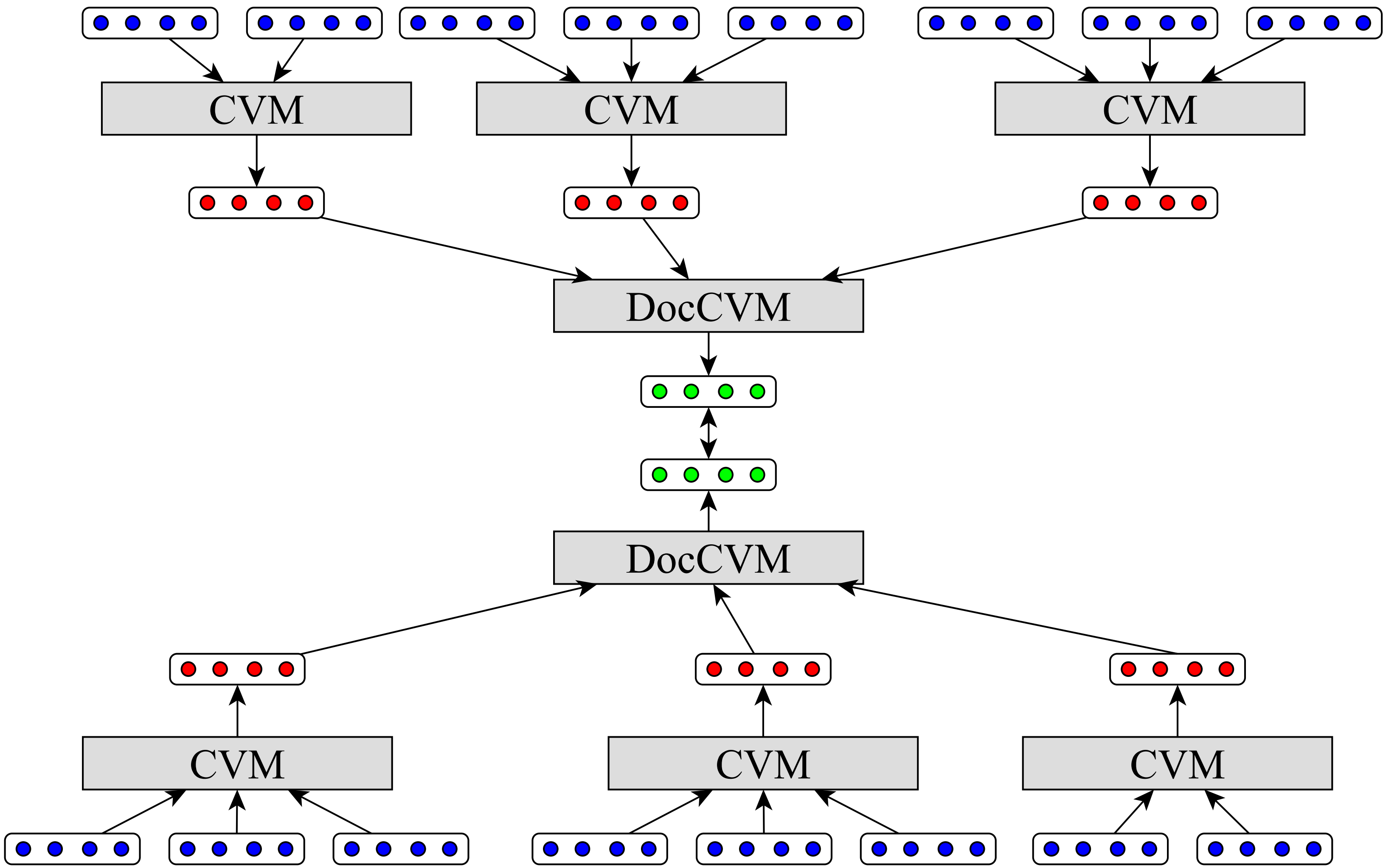
- Use noise samples:
  - $E_{hl}(x, y, n) = \max(0, m + E_{bi}(x, y) - E_{bi}(x, n))$
- Objective function:
  - $J(\theta) = \sum_{(x,y)\in C} \left(\sum_{i=1}^{k} E_{hl}(x, y, n_i) + \frac{\lambda}{2} \|\theta\|^2 \right)$

* AdaGrad, $m = d = 128, \lambda = 1, k \in \{1, 10, 50\}$

Compositional Vector Models (CVMs)

- Given a sentence $x = \{x_1, x_2, \ldots, x_n\}$
- add
  - $[\![x]\!] = \sum_{i=1}^{n} [\![x_i]\!]$
- bi(gram)*
  - $[\![x]\!] = \sum_{i=1}^{n} \tanh([\![x_{i-1}]\!] + [\![x_i]\!])$

Documents



Experiments

- Cross-lingual document classification
  - Embeddings: Europarl (en-fr, en-de)
  - Training/Test: RCV1/RCV2
- Multi-label document classification
  - Embeddings: TED
  - Training/Test: TED

Cross-lingual Document Classification

- Learn language-independent (?) word embeddings
- Train classifier on one language
- Test classifier on other language
- Representation of documents: average of representations of sentences
- Multi-class classifier trained using averaged perceptron; 15 classes

Cross-lingual Document Classification

Model ($d = 128$)	en $>$ de	de $>$ en
I-Matrix (Klementiev et al.)	77.6	71.1
add	86.4	74.7
add+	87.7	77.5
bi	86.1	79.0
bi+	**88.1**	**79.2**

X+: trained on 500k en-de pairs, and 500k en-fr pairs

Multi-Label Classification

- Learn word embeddings from 12 languages
  - Single training: learnt from single language pair
  - Joint training: learnt from all parallel sub-corpora
- doc models, i.e. doc/add and doc/bi
- Document representations used to train 12 classifiers (same as before; 15 classes)
- Baseline: MT system + NB classifier
  - "we do not expect to necessarily beat this system."

Multi-Label Classification

Setting	Languages										
	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Roman.	Russian	Turkish
en → L2											
MT System	0.429	0.465	0.518	0.526	0.514	0.505	0.445	0.470	0.493	0.432	0.409
ADD single	0.328	0.343	0.401	0.275	0.282	0.317	0.141	0.227	0.282	0.338	0.241
BI single	0.375	0.360	0.379	0.431	0.465	0.421	<u>0.435</u>	0.329	0.426	0.423	<u>0.481</u>
DOC/ADD single	<u>0.410</u>	0.424	0.383	<u>0.476</u>	<u>0.485</u>	0.264	0.402	0.354	0.418	0.448	0.452
DOC/BI single	0.389	<u>0.428</u>	0.416	0.445	0.473	0.219	0.403	0.400	<u>0.467</u>	0.421	0.457
DOC/ADD joint	0.392	0.405	0.443	0.447	0.475	<u>0.453</u>	0.394	<u>0.409</u>	0.446	<u>0.476</u>	0.417
DOC/BI joint	0.372	0.369	<u>0.451</u>	0.429	0.404	0.433	0.417	0.399	0.453	0.439	0.418
L2 → en											
MT System	0.448	0.469	0.486	0.358	0.481	0.463	0.460	0.374	0.486	0.404	0.441
ADD single	0.380	0.337	<u>0.446</u>	0.293	0.357	0.295	0.327	0.235	0.293	0.355	0.375
BI single	0.354	0.411	0.344	0.426	0.439	0.428	<u>0.443</u>	0.357	0.426	0.442	0.403
DOC/ADD single	<u>0.452</u>	<u>0.476</u>	0.422	0.464	<u>0.461</u>	0.251	0.400	0.338	0.407	<u>0.471</u>	0.435
DOC/BI single	0.406	0.442	0.365	<u>0.479</u>	0.460	0.235	0.393	0.380	0.426	0.467	<u>0.477</u>
DOC/ADD joint	0.396	0.388	0.399	0.415	<u>0.461</u>	<u>0.478</u>	0.352	<u>0.399</u>	0.412	0.343	0.343
DOC/BI joint	0.343	0.375	0.369	0.419	0.398	0.438	0.353	0.391	<u>0.430</u>	0.375	0.388

Multi-Label Classification: Linguistic Transfer

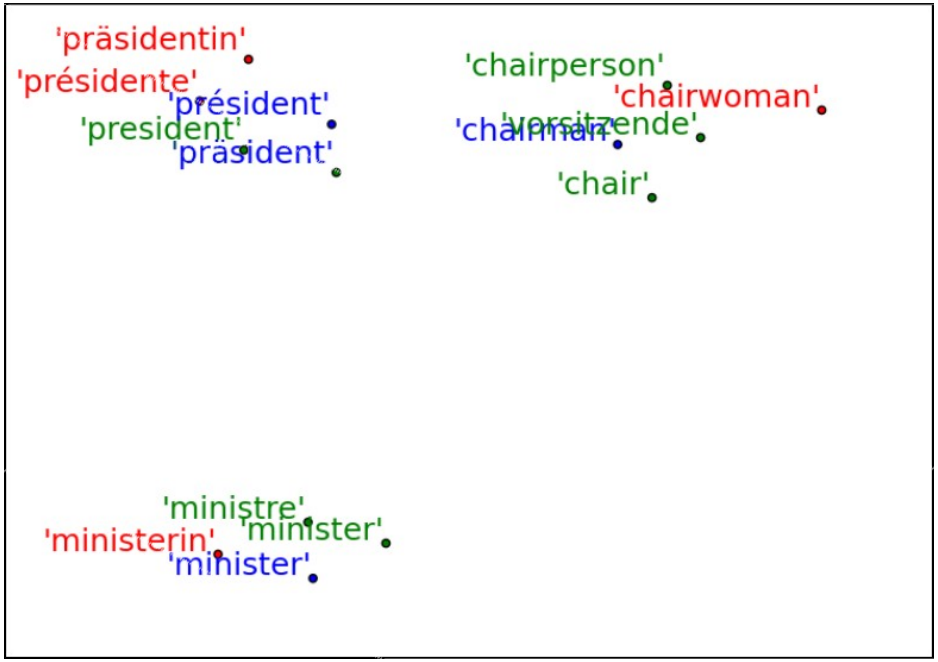
Training Language	Test Language										
	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Rom’n	Russian	Turkish
Arabic		0.378	0.436	0.432	0.444	0.438	0.389	0.425	0.420	0.446	0.397
German	0.368		0.474	0.460	0.464	0.440	0.375	0.417	0.447	0.458	0.443
Spanish	0.353	0.355		0.420	0.439	0.435	0.415	0.390	0.424	0.427	0.382
French	0.383	0.366	0.487		0.474	0.429	0.403	0.418	0.458	0.415	0.398
Italian	0.398	0.405	0.461	0.466		0.393	0.339	0.347	0.376	0.382	0.352
Dutch	0.377	0.354	0.463	0.464	0.460		0.405	0.386	0.415	0.407	0.395
Polish	0.359	0.386	0.449	0.444	0.430	0.441		0.401	0.434	0.398	0.408
Portuguese	0.391	0.392	0.476	0.447	0.486	0.458	0.403		0.457	0.431	0.431
Romanian	0.416	0.320	0.473	0.476	0.460	0.434	0.416	0.433		0.444	0.402
Russian	0.372	0.352	0.492	0.427	0.438	0.452	0.430	0.419	0.441		0.447
Turkish	0.376	0.352	0.479	0.433	0.427	0.423	0.439	0.367	0.434	0.411	

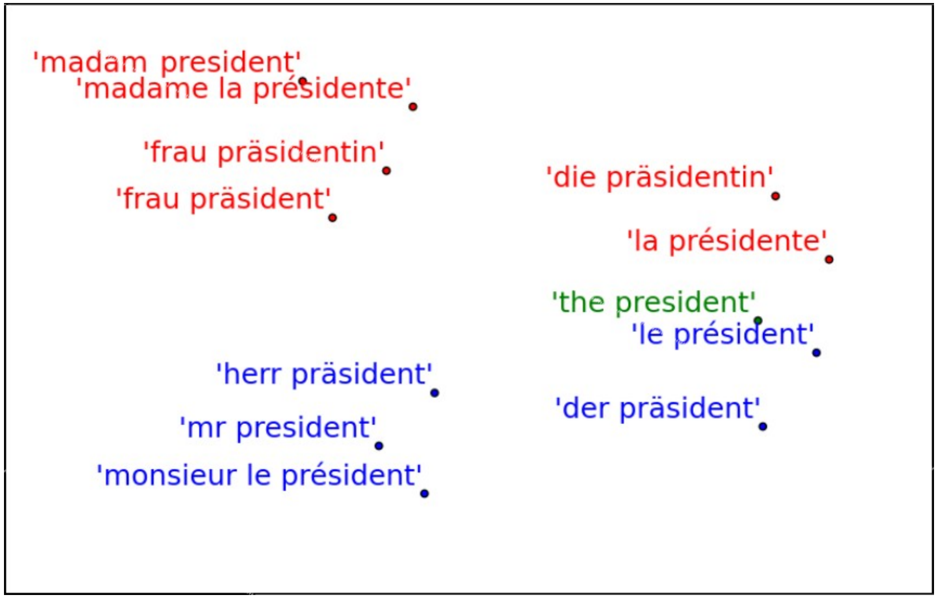
embeddings from doc/add joint model re-used to train classifiers on all non-English languages

Multi-Label Classification: Monolingual

Setting	Languages											
	English	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Roman.	Russian	Turkish
Raw Data NB	0.481	0.469	0.471	0.526	0.532	0.524	0.522	0.415	0.465	0.509	0.465	0.513
Senna	0.400											
Polyglot	0.382	0.416	0.270	0.418	0.361	0.332	0.228	0.323	0.194	0.300	0.402	0.295
single Setting												
DOC/ADD	0.462	0.422	0.429	0.394	0.481	0.458	0.252	0.385	0.363	0.431	0.471	0.435
DOC/BI	0.474	0.432	0.362	0.336	0.444	0.469	0.197	0.414	0.395	0.445	0.436	0.428
joint Setting												
DOC/ADD	0.475	0.371	0.386	0.472	0.451	0.398	0.439	0.304	0.394	0.453	0.402	0.441
DOC/BI	0.378	0.329	0.358	0.472	0.454	0.399	0.409	0.340	0.431	0.379	0.395	0.435

Projections





Conclusion

- Clever way of generating word embeddings
  - Not clear if these embeddings perform well as word embeddings per se
- Order of words in sentence?
- Order of sentences in document? (discourse)
- Order may not be that important?

Thank you!