Multilinguale Embeddings

Universität Heidelberg
Institut für Computerlinguistik
Embeddings
Katja Markert & Ines Rehbein
SoSe 19

Plan

- Einleitung und Motivation
- Modell
- Experimente
- Ergebnisse
- Fazit

Einleitung und Motivation

Einleitung

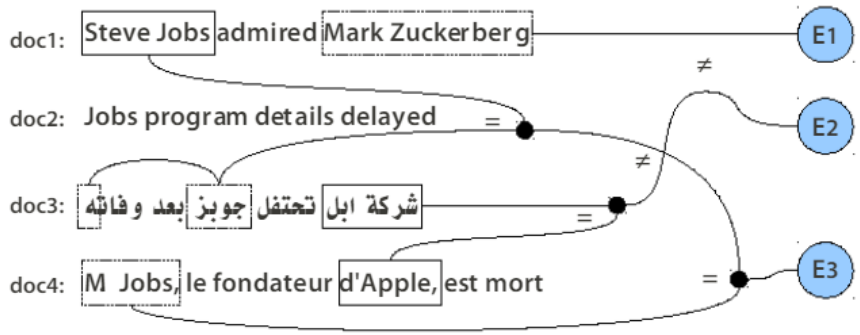
- Bisher:
  - Fokus auf einsprachigen Modellen für Vektorisierung in verschiedenen Tasks
  - Einbettung von unterschiedlichen Entitäten in denselben Raum (siehe Starspace)
  - Verwendung von gelabelter visueller Information (auch sprachübergreifend)
  - Multimodale Embeddings
  - sehr gute Ergebnisse für eine Sprache (Englisch)
- Problem:
  - Keine gute Methode für viele sprachübergreifende Anwendungen
  - In Sprachen außer Englisch waren die Ergebnisse nicht überzeugend
- Lösung:
  - Multilinguale Embeddings

Motivation

- Embeddings, die sowohl monolingual, als auch bilingual gut sind
  - Vor Luong et al. (2015) waren viele Ansätze nur in biligualen Anwendungen gut

Motivation

- Andere mögliche Anwendungen von multilingualen Embeddings:
  - Maschinelle Übersetzung (Brown et al., 1993)
  - Noun bracketing (Yarowsky and Ngai, 2001)
    - (woman (aid worker)) → right-bracketing interpretation
    - ((copper alloy) rod) → left-bracketing interpretation
  - Entity Clustering (Green et al., 2012)



  - Bilinguale NER (Named Entity Recognition) (Wang et al., 2013):
    - Jim bought 300 shares of Acme Corp. in 2006. →
    - $[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[2006]_{\text{Time}}$.

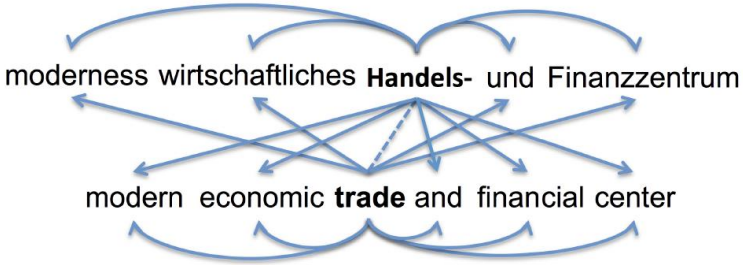
ME – Herangehensweisen

- Bilinguales Mapping
  - Embeddings in 2 Sprachen einzeln lernen
  - Ein Mapping zwischen 2 Sprachen lernen
- Monolinguale Adaptation
  - Embeddings in einer reichen Sprache 1 lernen
  - Mithilfe eines bilingualen Constraints Embeddings in Sprache 2 so lernen, dass gleiche Wörter nah aneinander eingebettet sind
- Bilinguales Training
  - Embeddings in Sprache 1 und 2 gleichzeitig lernen
  - Ansatz von Luong und Kollegen

Modell

Modell

- SkipGram mit Negative Sampling
- $\alpha(Mono_1 + Mono_2) + \beta Bi$, wo
  - $Mono_x$ das Modell für Vorhersagen innerhalb Sprache x ist
  - Bi das Modell für Vorhersagen zwischen 2 Sprachen



- SkipGram so erweitert, dass Wörter sprachübergreifend vorhergesagt werden können
- Richtungen des Lernprozesses:
  - $l_1 \rightarrow l_1$, innerhalb Sprache 1
  - $l_2 \rightarrow l_2$, innerhalb Sprache 2
  - $l_1 \rightarrow l_2$, mit einem Embedding aus Sprache 1 Wörter in Sprache 2
  - $l_2 \rightarrow l_1$, mit einem Embedding aus Sprache 2 Wörter in Sprache 1

Modell – Alignierungen

- BiSkip-UnsupAlign:
  - Berkeley aligner (Liang et al., 2006)
  - Alignierungen werden unüberwacht gelernt
- BiSkip-MonoAlign:
  - Annahme:
    - Wörter sind im Parallelkorpus monoton aligniert
  - Alignierung:
    - $i \cdot \frac{T}{S}$
      - i – Position des Wortes im Quellsatz
      - T – Länge des Zielsatzes
      - S – Länge des Quellsatzes

Experimente

Experimente – Korpus

- Europarl-Korpus (v7, (Koehn, **2005**)):
- 1.9 Millionen paralleler Sätze (de-en)
- Protokolle der Sitzungen des Europäischen Parlamentes
  - umfasst **aktuell** 21 offizielle Sprachen
- Sätze schon durch Algorithmus von Gale & Church (1993) aligniert
- Zahlen mit x Ziffern durch x Nullen ersetzt $(1999 \rightarrow 0000)$
  - alle gleich langen Zahlen erhalten gleiche Bedeutung
- Wörter mit freq < 5 werden zu <unk>
- Vokabulargröße:
  - Englisch: 40K
  - Deutsch: 95K

Experimente – Modellparameter

- Learning rate = α = 0.025:
  - $\theta_{new} = \theta_{old} - \alpha \nabla_{\theta} J(\theta)$
- Negative Sampling:
  - 30 negative pro 1 positives Sample
- Kontextfenstergröße = 5 Wörter
- Subsampling rate = t = 10^{-4}:
  - $P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$
- $\alpha(Mono_1 + Mono_2) + \beta Bi$
  - $\alpha = 1$
  - $\beta = 4$

Experimente – Evaluation Tasks

- Monolinguale Tasks:
  - Wortähnlichkeit
- Bilinguale Tasks
  - CLDC – Cross-lingual Document Classification

Word Similarity

Evaluation Tasks – Word Similarity

- Man untersucht semantische Qualität von Embeddings
- Spearman‘s Rank Corellation Coefficient zwischen Ähnlichkeitswerten von 2 Embeddings und menschlichen Annotatoren
- Dafür verwendete Datasets:
  - WordSim (353 Paare)
  - RG (65 Paare)
    - MC (30 Paare) (subset)
    - Korrelation von 0.968 zwischen MC und RG
  - SCWS (1762 Paare)
  - RW (2034 Paare)

RG-65

Info:

- Dataset von Rubenstein & Goodenough (1965)
- 65 Wortpaare, nur Substantive
- Ähnlichkeit von jedem Paar wird von 0 bis 4 bewertet
- Gerankt von 2 Gruppen von Studenten
- Die Similarity-Werte sind Durchschnittswerte von 51 Bewertungen
- Für Europarl als Testset geignet?

Paare (Beispiele):

Wort 1	Wort 2	Bewertung
gem	jewel	3.94
midday	noon	3.92
bird	cock	2.63
brother	lad	2.41
bird	woodland	1.24
food	rooster	1.09
shore	voyage	1.22
monk	slave	0.57
automobile	wizard	0.11

SCWS*

Info

- Modifizierte Version von Stanford's contextual word similarities dataset
- 1762 Wortpaare, verschiedene POS
- Ähnlichkeit von 0 bis 10
- Im Original-Dataset ist der umgebende Kontext wichtig!
  - financial bank vs river bank
  - Solche Paare sind im SCWS* exkludiert
- Was fällt auf im Gegensatz zu RG?

Paare (Beispiele)

Wort 1	Wort 2	Bewertung
abuse	persecution	5.7
Brazil	nut	1.1
minister	foreign	3.15
ministry	building	2.3
mock	tease	7.5
mole	spy	4.6
money	bank	4.2
stock	market	4.7
Arafat	terror	2.72

RW

Info

- Rare words dataset von (Luong et al., 2013)
- 2034 Paare
- Ähnlichkeit von 0 bis 10
- Erstellungsprozess:
  - Finde eine Liste x von seltenen Wörter
    - 5 frequency bins
  - Bilde Paare W1 (RW) – 2*W2 (aus dem WordNet-Synset von W1)
  - Lass Menschen annotieren

Paare (Beispiele)

Wort 1	Wort 2
untracked	inaccessible
unflagging	constant
unprecedented	new
apocalyptical	prophetic
organismal	system
diagonal	line
obtainment	acquiring
discernment	knowing
confinement	restraint

Datasets – Zusammenfassung

	pairs	type	raters	scale	Complex words	
					token	type
WS353	353	437	13-16	0-10	24	17
MC	30	39	38	0-4	0	0
RG	65	48	51	0-4	0	0
SCWS*	1762	1703	10	0-10	190	113
RW (new)	2034	2951	10	0-10	987	686

Table 3: **Word similarity datasets** and their statistics: number of pairs/raters/type counts as well as rating scales. The number of complex words are shown as well (both type and token counts). RW denotes our new rare word dataset.

Datasets – Zusammenfassung

	All words	Complex words
WS353	0 | 0 / 9 / 87 / 341	0 | 0 / 1 / 6 / 10
MC	0 | 0 / 1 / 17 / 21	0 | 0 / 0 / 0 / 0
RG	0 | 0 / 4 / 22 / 22	0 | 0 / 0 / 0 / 0
SCWS*	26 | 2 / 140 / 472 / 1063	8 | 2 / 22 / 44 / 45
RW	801 | 41 / 676 / 719 / 714	621 | 34 / 311 / 238 / 103

Table 4: **Word distribution by frequencies** – distinct words in each dataset are grouped based on frequencies and counts are reported for the following bins : unknown | [1, 100] / [101, 1000] / [1001, 10000] / [10001, ∞). We report counts for all words in each dataset as well as complex ones.

Cross-lingual Document Classification

- Zeigt die Qualität der bilingualen Komponente
- Training mit 1000 Dokumenten:
  - Man trainiert zunächst, die Docs in Sprache 1 zu klassifizieren:
    - Averaged perceptron algorithm
  - Wendet danach das Modell (Gewichtsmatrizen W_n der Hidden Layers) auf Sprache 2 an, sodass $x_{BiSkip} \cdot W_n \rightarrow label$
  - $doc_{repr} = \frac{1}{len(doc)} \sum_{i=1}^{len(doc)} inverteddocfreq\ (w_i, doc) \cdot \mathrm{w}_i$
  - $inverteddocfreq(w_i, doc) = log \frac{N(D)}{N(D, x_i)}$
- Testen auf 5000 labeled-RCV (Reuters Corpus Volume) Dokumenten
- Prämisse:
  - Wörter sind im gleichen Raum eingebettet → Docs haben ähnliche Repräsentationen

Ergebnisse

Ergebnisse

Models	Dim	Data	Word Similarity						CLDC	
			de	en					en→de	de→en
			WS353	WS353	MC	RG	SCWS	RW		
Existing best models										
I-Matrix	40	Europarl+RCV	23.8	13.2	18.6	16.4	19.0	07.3	77.6	71.1
BilBOWA	40	Europarl+RCV	-	-	-	-	-	-	86.5	75.0
DWA	40	Europarl	-	-	-	-	-	-	83.1	75.4
BAE-cr	40	Europarl+RCV	34.6	39.8	32.1	24.8	29.3	20.5	**91.8**	74.2
CVM-Add	128	Europarl	28.3	19.8	21.5	24.0	28.9	13.6	86.4	74.7
Our BiSkip models										
MonoAlign	40	Europarl	43.8	41.0	33.9	32.2	39.5	24.4	86.4	75.6
	128	Europarl	45.9	46.0	30.4	27.1	**43.4**	**25.3**	89.5	78.4
UnsupAlign	40	Europarl	43.0	40.2	31.7	32.1	37.6	23.1	87.6	77.8
	128	Europarl	45.5	45.8	36.6	32.3	42.3	*24.6*	88.9	77.4
	256	Europarl	*46.7*	*47.3*	*37.9*	**35.1**	43.2	24.5	88.4	**80.3**
	512	Europarl	**47.4**	**49.3**	**45.7**	**35.1**	**43.4**	24.0	*90.7*	*80.0*

Ergebnisse – Nearest neighbor words

	january		**microsoft**		**distinctive**		
	en	*de*	*en*	*de*	*en*	*de*	*gloss*
BiSkip	january	januar	microsoft	microsoft	distinctive	unverwechselbare	distinctive
	july	februar	ibm	ibm	character	darbietet	presents
	december	juli	linux	walt	features	eigenheit	peculiarity
	october	dezember	ipad	mci	individualist	unschtzbarer	invaluable
	march	november	blockbuster	linux	patrimony	charakteristische	characteristic
	february	jahres	doubleclick	kurier	diplomacies	identittsstiftende	identity
	april	oktober	yahoo	setanta	splendour	christlich-jdischen	christian-jewish
	november	april	rupert	yahoo	vocations	identittsfindung	identity-making
	september	august	alcatel	warner	multi-faith	zivilisationsprojekt	civilization project
	august	juni	siemens	rhne-poulenc	characteristics	ost-west-konflikt	east-west conflict
Autoencoder	january	januar	microsoft	microsoft	distinctive	rang	rank
	march	mrz	cds	cds	asset	wiederentdeckung	rediscovery
	october	oktober	insider	warner	characteristic	echtes	real
	july	juli	ibm	tageszeitungen	distinct	bestimmend	determining
	december	dezember	acquisitions	ibm	predominant	typischen	typical
	1999	jahres	shareholding	telekommun*	characterise	bereichert	enriched
	june	juni	warner	handelskammer	derive	sichtbaren	visible
	month	1999	online	exchange	par	band	band
	year	jahr	shareholder	veranstalter	unique	ausgeprgte	pronounced
	september	jahresende	otc	geschftsfhrer	embraces	vorherrschende	predominant

Table 2: **Nearest neighbor words** – shown are the top 10 nearest English (en) and German (de) words for each of the following words in the list {*january, microsoft, distinctive*} as measured by the Euclidean distances given a set of embeddings. We compare our learned vectors (BiSkip-UnsupAlign, $d = 128$) with those produced by the autoencoder model (Chandar A P et al., 2014). For the word *distinctive*, we provide Google Translate glosses for German words. The word *telekommunikationsunternehmen* is truncated into telekommun*.

Ergebnisse – Wortvisualisierungen

Figure 2: Bilingual Embeddings of German and English words about banking and financial

Fazit

- Vor Luong et al. (2015):
  - Modelle haben gute Performanz nur in bilingualen Anwendungen
  - Clustering von Wortrepräsentationen nur bilingual
  - Monolingual – schlechte Ergebnisse
- Luong et al. (2015):
  - Bi- und monolinguale Komponente gleichgewichtet
  - State-of-the-art-Akkuratheit von 80.3% in CLDC (crosslingual document classification) (Deutsch zu Englisch)
  - Sehr gutes Ergebnis von 90.7% in CLDC (Englisch zu Deutsch)
  - übertreffen vorherige beste Embeddings in monolingualer Wortähnlichkeit mit großem Abstand

Vielen Dank für Ihre Aufmerksamkeit!