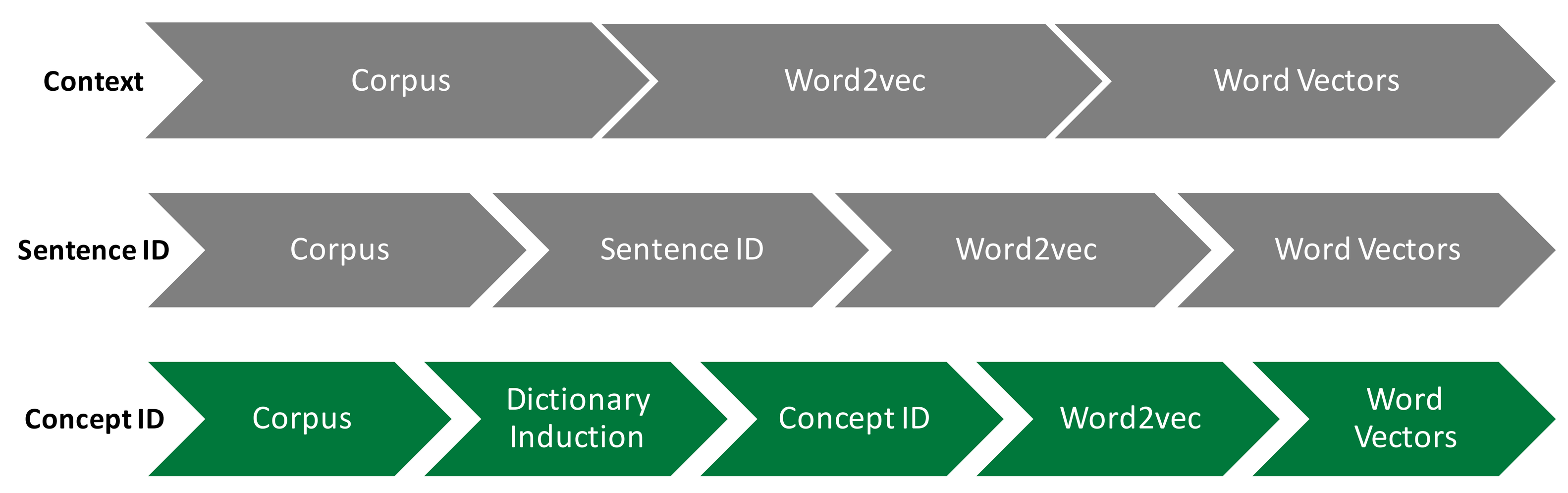
Embedding Learning Through Multilingual Concept Induction

Philipp Dufter[1], Mengjie Zhao[2], Martin Schmitt[1], Alexander Fraser[1], Hinrich Schütze[1]

[1] Center for Information and Language Processing (CIS) LMU Munich, Germany
[2] École Polytechnique Fédérale de Lausanne, Switzerland

{philipp,martin,fraser}@cis.lmu.de, mengjie.zhao@epfl.ch

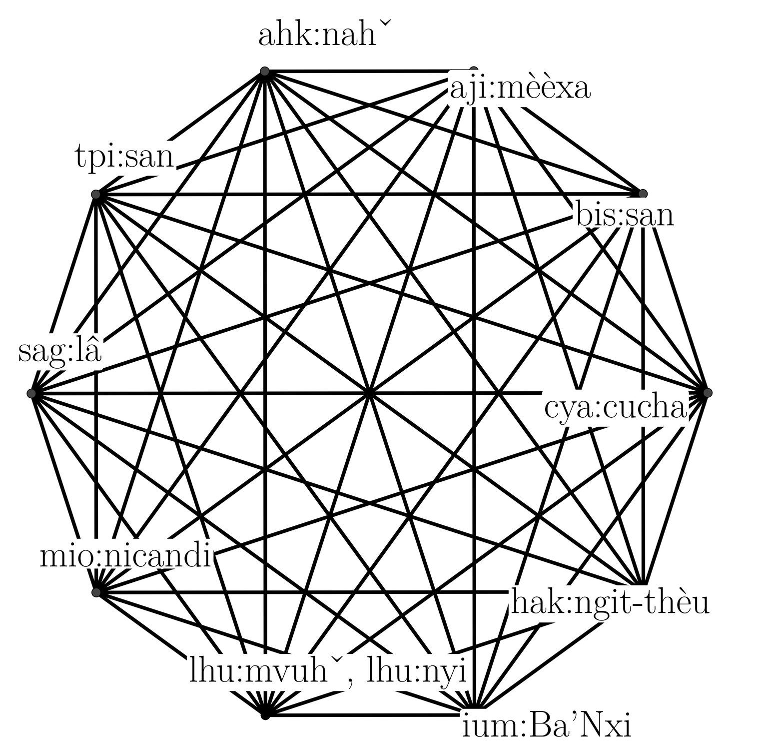
Motivation



Objective: **Multilingual wordspace with 1000s (low-resource) languages.**

New feature: **concept-IDs**.

The concept "sun"




```
[...]
ell4:ήλιου aby0:eweka aby0:ewekanu aby0:ewekaini wrs0:okombava
wrs0:okomba wrs0:okombam cbs0:badi bjr0:súani bjr0:Súani
bjr0:afu deu2:Sonne engb:the engb:sun fra2:soleil hwc1:sun
hwc1:da med0:ant nob3:solen nob3:Solen spa1:sol sua0:kolkha
sua0:a tir0:ጸሓይ slv0:sonce slv0:sonca slv0:Sonce snn2:xɨcaj
xav0:bâtâ arb1:ٱلشَّمْسِ mcp0:jwɔ́wʉ́d mcp0:yásə mco0:xɨɨ mcd0:fari
men0:foloi mdy0:ዓቢ mdy0:ዓቢፖዲ mcf0:ushë tir1:ጸሓይ tur4:güneş
tur4:güneşe tur4:Güneş zho3:太陽 zho6:日頭
[...]
```

	roundtrip translation for "sun" S1 (μ)	R1 (μ)	S4 (μ)	S16 (μ)
S-ID	61	61	83	99
CLIQUE	**85**	**85**	**98**	**100**

ISO-639-3	Language	Family	Region (example)
lhu	Lahu	Sino-Tibetan	Thailand
ahk	Akha	Sino-Tibetan	China
hak	Hakka Chinese	Chinese	China
ium	Iu Mien	Hmong-Mien	Laos
tpi	Tok Pisin	Creole	Papua New Guinea
mio	Pinotepa Mixtec	Oto-Manguean	Oaxaca
cya	Highland Chatino	Oto-Manguean	Oaxaca
bis	Bislama	Creole	Vanuatu
aji	Ajië	Austronesian	New Caledonia
sag	Sango	Creole	Central Africa

Contributions

1. Novel embedding learning method: **concept-based embedding learning**.
2. New word-/character-level **dictionary and concept induction** methods.
3. Word translation and sentiment analysis **across 1259 languages**.

Conclusions

1. **Concept-based methods outperform** previous approaches.
2. New roundtrip evaluation is an **excellent wordspace quality indicator**.
3. **Character-level is better than word-level** for sentiment classification.

Data

- Parallel Bible Corpus [Mayer and Cysouw, 2014]: 7958 verses, 1259 languages, 1664 editions.

English King James Version	German Elberfelder 1905	Spanish Americas
And he said , Do it the second time . And they did it the second time . . . And he said , Do it the third time . . . third time .	Und er sprach : Füllet vier Eimer mit Wasser , und gießet es auf das Brandopfer und auf das Holz . Und er sprach : Tut es zum zweiten Male ! Und sie taten es zum zweiten Male . . . Und er sprach : Tut es zum dritten Male ! Und sie taten es zum dritten Male .	Y dijo : Llenad cuatro cántaros de agua y derramadla sobre el holocausto y sobre la leña . Después dijo : Hacedlo por segunda vez ; y lo hicieron por segunda vez . . . Y añadió : Hacedlo por tercera vez ; y lo hicieron por tercera vez .

Dictionary Induction

- Creating $\mathcal{O}(n^2)$ dictionaries for n languages not scalable
  $\Rightarrow$ selection of $p = 10$ pivot languages
- Computing $p(p-1)/2 + pn$ dictionaries (intra-pivot and pivot-target) by aggregating fast-align alignments.

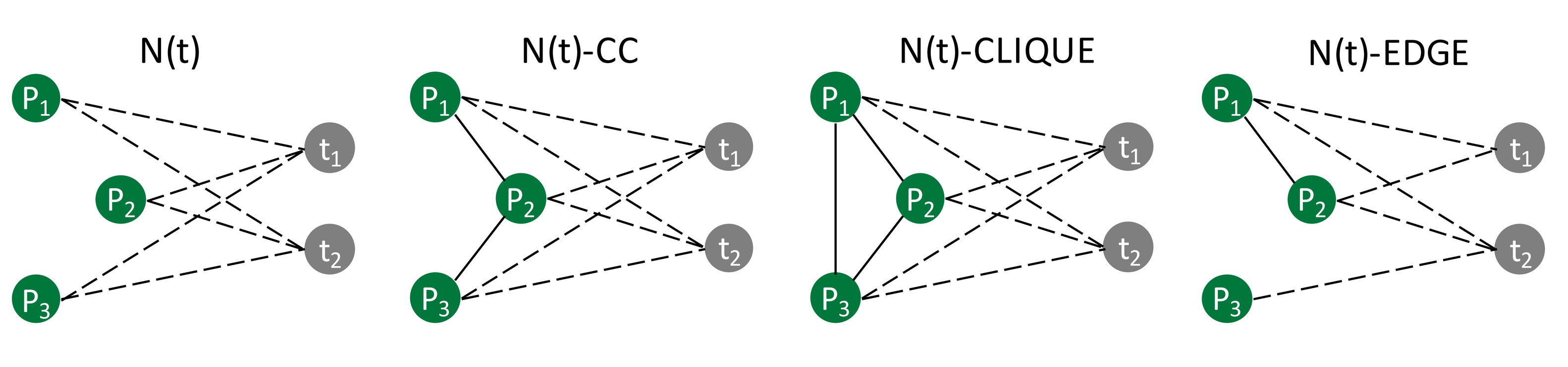
Concept Induction

1) CLIQUE Projection

- In ideal world: concepts correspond to cliques in the multilingual dictionary graph.
- In real-world: identify quasi-cliques to accommodate noise.
- Identify concepts in pivot dictionary graph and project them onto target languages.

2) Target Neighborhoods N(t)

- Define target neighborhood for a word t by $N(t) := \{w \in V_p \,|\, (w, t) \in E_D\}$.
- Two words t_1 and t_2 are in the same concept if $N(t_1) = N(t_2)$.



Baselines

- Non-wordspace based: **RTSIMPLE**. Roundtrip translation directly on the dictionary graph.
- Context-based wordspaces: bag-of-words (**BOW**) [Vulić et al., 2015], sentence-ID (**S-ID**) [Levy et al., 2017].
- Alternative concept-identification method: **SAMPLE** [Lardilleux et al., 2009]. Sampling based approach for identifying concepts in a sentence-aligned corpus. Projection step same as for CLIQUE.

S-ID vs. C-ID

``...and gather his wheat into the garner ; but he will burn up the chaff with unquenchable fire .''

	q	$\Rightarrow$	$I_e(q)$	$\Rightarrow$	$T_e(q)$
S-ID	burn	$\Rightarrow$	gogoro	$\Rightarrow$	harvest labourers fowls wheat sow gather tares gnashing
		$\Rightarrow$	gbi	$\Rightarrow$	fire burned brimstone hell burn burning smoke Sodom flame goats gnashing offerings vial devour wheat perdition
CLIQUE	burn	$\Rightarrow$	agbi	$\Rightarrow$	burn burned burning burning furnace fire flame warming unquenchable
		$\Rightarrow$	gbi	$\Rightarrow$	burning burn fire furnace burned unquenchable flame warming smoke lampstands candle lamps extinguished lamp pool Hell

	roundtrip translation for "burn" S1 (μ)	R1 (μ)	S4 (μ)	S16 (μ)
S-ID	0.06	0.18	0.14	0.72
CLIQUE	**0.72**	**0.92**	**0.95**	**1.00**

- Roundtrip translation fails for query `burn` in language sag.
- Reason: `sag:gogoro` occurs only in the context of `burning`
- Nearest neighbours of `sag:gbi` within eng reflect the context of `burn` instead of semantically related words

WORD vs. CHAR

```
[ksw] ဒီးတၢ်ကမၣ်လၢအဟ်လၢယလိၤခဲကနံၣ်အံၤ⋆,
      ⋆ထုးပှၢ်အှၣ်အသးတနၢ်ဘၣ်⋆.
[cso] Hi³⋆sa³jun³⋆lɨ́¹³⋆ma³tson²⋆tsú²⋆
      lɨ³ua³⋆cáun²⋆tso³⋆ñí¹⋆hná¹⋆nɨ́²⋆.
[eng] Neither⋆can⋆they⋆prove⋆the⋆things⋆
      whereof⋆they⋆now⋆accuse⋆me⋆.
```

Many hard-to-tokenize and morphologically rich languages
$\Rightarrow$ create 2 versions per edition: **WORD** and **CHAR**
WORD (whitespace tok.):
`[Neither, can, they, prove, the, things, ...`
CHAR (overlapping byte ngrams):
`[Neit, eith, ithe, ther, her@, er@c, r@ca, @can, ...`

Adjustments:

- χ^2 **based dictionary:** iteratively select word pair with highest χ^2. Then remove cooccurrence of this pair.
- **Query selection RTT:** find ngrams for each query word which correspond uniquely to this word.
- **Groundtruth in RTT:** select words based on degree of correspondance between ngram and query word.
- **Sentiment analysis:** consider only ngrams which have been aligned.

Embedding Learning

- Word2vec skipgram model with mostly default hyperparameters.

BOW
- ...
- engw:. tpi:' mio:cuu mio:, engw:me engw:" mio:ri engw:. engw:You mio:chata mio:yuhu engw:years mio:Israel engw:They mio:ndo mio:yoo mio:sacahnu [...] tpi:animal engw:sky engw:and mio:vati
- sag:so ahk:, ahk:tsawˇ sag:so ahk:lawˇ-eu̦ ahk:Iˇsa̦le̦laˆ ahk:dahˇ [...] ahk:yaŵ sag:Nilaa ahk:ngeh̦ engw:. ahk:a̦ ahk:miˇ ahk:je̦ ahk:guiˇ ahk:bawˇ ahk:du̦ engw:the ahk:m̦ engw:sky sag:ni engw:sun ahk:tseˇ ahk:ma sag:Nzapa ahk:mâ ahk:za̦
- ...

S-ID
- ...
- v4013 mio:Israel
- v4013 mio:ñivi
- v4013 mio:cati
- v4013 mio:tichi
- v4013 mio:nicandi
- v4014 mio:luhu
- v4014 mio:Cuhva
- ...

Concept-Based
- ...
- mdy0:ዓቢፖዲ guj0:સૂર્ય wca0:pili che0:Малх tgl3:ang C:30771:bis:san aeu0:nlma kor2:haywa xav0:bâdâ mmn0:na pol5:słońce
- ces1:Slunce blw0:init uig2:күн mio:nicandi mqb0:dav C:30771:bis:san engk:the eng7:the lug1:njuba qvn0:intis sld0:wɩpomí
- nma0:leimik amh2:ፀሐይ hmo0:dina ame0:atsne' zpm0:ngbidz C:52750:bis:san ium3:Mba'hnoi daa0:pat dop0:ilim xav0:Bâdâ hun5:nap
- ...

Evaluation

Roundtrip translation

q	$\Rightarrow$	$I_e(q)$	$\Rightarrow$	$T_e(q)$	$S1$	$R1$	$S4$	$S16$
woman	$\Rightarrow$	mujer	$\Rightarrow$	wife woman women widows daughters daughter marry married				
	$\Rightarrow$	esposa	$\Rightarrow$	marry wife woman married marriage virgin daughters bridegroom	$+0$	$+0$	$+1$	$+1$

- 70 English queries taken from a list of universal words by Swadesh (1946).
- Strict and relaxed groundtruth: $G_s(q) = \{q\}$, $G_r(q)$ contains words with the same lemma as q.
- Accuracy computed by $1/|E| \sum_{e \in E} \min\{1, |T_e(q) \cap G_i(q)|\}$, aggregated over queries.
- $(i, |I_e(q)|, |T_e(q)|)$ is varied as follows: "S1"$(s, 1, 1)$, "S4" $(s, 2, 2)$, "S16" $(s, 2, 8)$, and "R1" $(r, 1, 1)$.

Sentiment Analysis

``Now is come salvation ... the power of his Christ: for the accuser ... cast down, which accused them before our God ...''

- Creation of English silver standard using the Vader-Classifier in combination with manual annotations
- Unclear sentiment: $\Rightarrow$ 2 tasks: "contains positive sentiment" and "contains negative sentiment"
- Assumption: each verse has unchanged sentiment across languages.
- Linear SVMs for classification and report of average F_1 scores (across languages)

Results

		roundtrip translation WORD S1		R1		S4		S16			CHAR S1		R1		S4		S16			sentiment analysis WORD		CHAR	
		μ	Md	μ	Md	μ	Md	μ	Md	N	μ	Md	μ	Md	μ	Md	μ	Md	N	pos	neg	pos	neg
1	RTSIMPLE	33	24	37	36					67	24	13	32	21					70				
2	BOW	7	5	8	7	13	12	26	28	69	3	2	3	2	5	4	10	11	70	33	81	13	83
3	S-ID	46	46	52	55	63	76	79	91	65	9	5	9	5	14	9	25	22	70	79	88	65	86
4	SAMPLE	33	23	43	42	54	59	82	96	65	53	**59**	59	**72**	67	85	79	99	58	82	89	77	89
5	CLIQUE	43	36	59	63	67	77	93	99	69	42	46	48	55	60	76	73	98	53	84	89	69	88
6	N(t)	**54**	**59**	**61**	69	**80**	**87**	**94**	**100**	69	50	53	54	59	73	82	90	99	66	82	89	**87**	**90**
7	N(t)-CC	52	56	59	66	77	86	93	99	69	40	45	42	48	58	69	75	95	57	80	88	68	86
8	N(t)-CLIQUE	11	0	11	0	16	0	22	0	18	39	45	41	47	58	74	76	94	56	22	84	61	84
9	N(t)-EDGE	35	30	43	36	56	55	87	94	69	39	29	49	52	64	78	88	**100**	63	84	**90**	84	89

We gratefully **acknowledge** funding from the European Research Council (grants 740516 & 640550) and through a Zentrum Digitalisierung.Bayern fellowship awarded to the first author. We are indebted to Michael Cysouw for making PBC available to us.

Presented at ACL 2018, Melbourne.