Enriching word vectors with subword information

(paper review)

Khoklin I., Kovrigin A., Shibaev E.

Motivation

Why embedding words considering each word as a **whole** is not the best idea?

Motivation: morphologically rich languages

- Finnish has **15 cases for nouns**.
- In French and Spanish most verbs have more than **40 different inflected forms**.

The verb *parler* "to speak", in French orthography and IPA transcription

	Indicative				Subjunctive		Conditional	Imperative
	Present	**Simple past**	**Imperfect**	**Simple future**	**Present**	**Imperfect**	**Present**	**Present**
je	parl-e /paʁl/	parl-ai /paʁle/	parl-ais /paʁlɛ/	parl-erai /paʁləʁe/	parl-e /paʁl/	parl-asse /paʁlas/	parl-erais /paʁləʁɛ/	
tu	parl-es /paʁl/	parl-as /paʁla/	parl-ais /paʁlɛ/	parl-eras /paʁləʁa/	parl-es /paʁl/	parl-asses /paʁlas/	parl-erais /paʁləʁɛ/	parl-e /paʁl/
il	parl-e /paʁl/	parl-a /paʁla/	parl-ait /paʁlɛ/	parl-era /paʁləʁa/	parl-e /paʁl/	parl-ât /paʁlɑ/	parl-erait /paʁləʁɛ/	
nous	parl-ons /paʁlɔ̃/	parl-âmes /paʁlɑm/	parl-ions /paʁljɔ̃/	parl-erons /paʁləʁɔ̃/	parl-ions /paʁljɔ̃/	parl-assions /paʁlasjɔ̃/	parl-erions /paʁləʁjɔ̃/	parl-ons /paʁlɔ̃/
vous	parl-ez /paʁle/	parl-âtes /paʁlɑt/	parl-iez /paʁlje/	parl-erez /paʁləʁe/	parl-iez /paʁlje/	parl-assiez /paʁlasje/	parl-eriez /paʁləʁje/	parl-ez /paʁle/
ils	parl-ent /paʁl/	parl-èrent /paʁlɛːʁ/	parl-aient /paʁlɛ/	parl-eront /paʁləʁɔ̃/	parl-ent /paʁl/	parl-assent /paʁlas/	parl-eraient /paʁləʁɛ/	

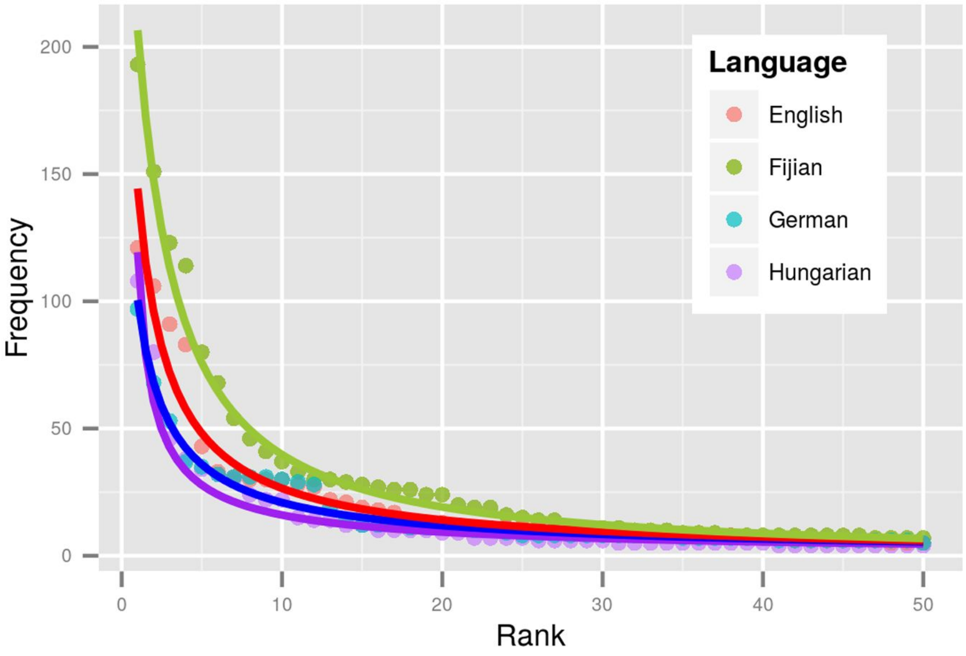
The conjugations of a verb "to speak" in French.

talo 'house'	singular	plural
nominative	talo	talot
genitive	talon	talojen
partitive	taloa	taloja
inessive	talossa	taloissa
elative	talosta	taloista
illative	taloon	taloihin
adessive	talolla	taloilla
ablative	talolta	taloilta
allative	talolle	taloille
essive	talona	taloina
translative	taloksi	taloiksi
instructive	-	taloin
abessive	talotta	taloitta
comitative	-	taloine+POS

All declinations of a word "house" in Finnish

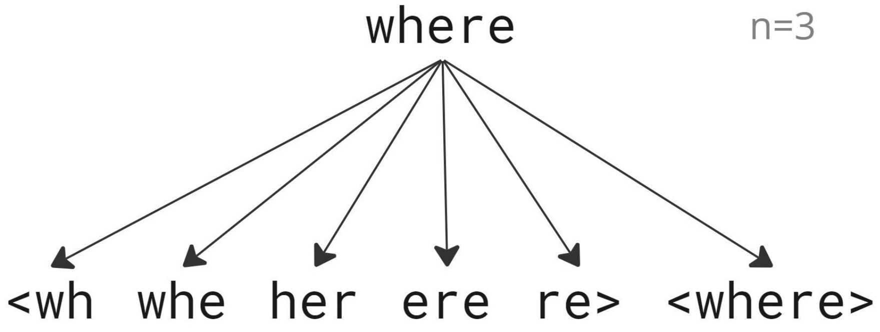
Motivation: rare words & amount of data for training

- Languages have **big tails of rare words**.
- Hence, “whole word embeddings” require an **immense amount of data** to learn properly and, for rare words, the **embeddings end up being bad**.



Proposal: Subword model

- Proposal: **represent each word by character n-grams**
- Word embedding is a **sum of n-gram embeddings**
- Technicalities:
  - Boundary symbols < and > added
  - The whole word is added as a special sequence
  - In practice, all n-grams where 3 <= n <= 6 are taken



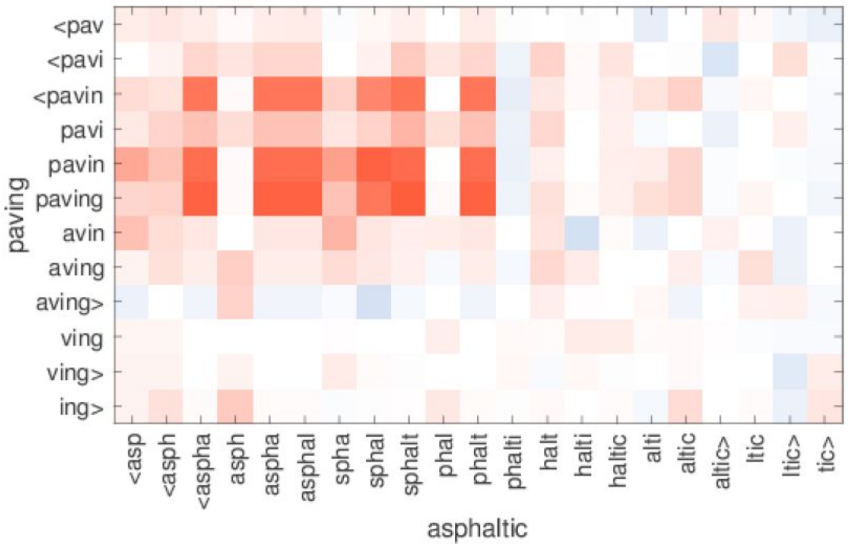
Qualitative analysis: Subword importance

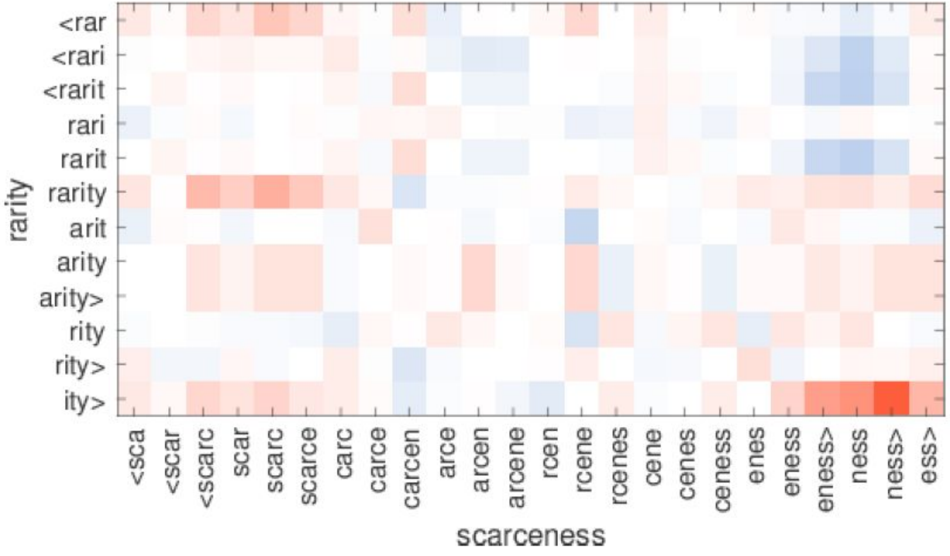
- Analyze learned embeddings
- **Rank n-grams by importance**
  - Remove n-gram from the sum
  - Compute cosine similarity between word embedding and word embedding without the n-gram
  - **Lower similarity → higher importance**

	word	n-grams		
DE	autofahrer	fahr	fahrer	auto
	freundeskreis	kreis	kreis>	<freun
	grundwort	wort	wort>	grund
	sprachschule	schul	hschul	sprach
	tageslicht	licht	gesl	tages
EN	anarchy	chy	<anar	narchy
	monarchy	monarc	chy	<monar
	kindness	ness>	ness	kind
	politeness	polite	ness>	eness>
	unlucky	<un	cky>	nlucky
	lifetime	life	<life	time
	starfish	fish	fish>	star
	submarine	marine	sub	marin
	transform	trans	<trans	form
FR	finirais	ais>	nir	fini
	finissent	ent>	finiss	<finis
	finissions	ions>	finiss	sions>

Qualitative analysis: Subword similarity

- Discover which n-grams are considered similar
- Calculate **cosine similarity** between subword embeddings





Results. Human similarity judgement

The authors compared word similarity through human judgment and cosine similarity of generated embeddings across various languages. They also distinguished between common and rare English words in their datasets.

Their method was compared with CBOW and Skip-gram baselines, showing superior performance on all datasets, with the exception of the English WS353 dataset. Notably, the proposed model's computation of vectors for out-of-vocabulary words (sisg) consistently surpassed the alternative approach of not computing these vectors (sisg-).

		sg	cbow	sisg-	sisg
AR	WS353	51	52	54	**55**
DE	GUR350	61	62	64	**70**
	GUR65	78	78	**81**	**81**
	ZG222	35	38	41	**44**
EN	RW	43	43	46	**47**
	WS353	72	**73**	71	71
ES	WS353	57	58	58	**59**
FR	RG65	70	69	**75**	**75**
RO	WS353	48	52	51	**54**
RU	HJ	59	60	60	**66**

There are two ways to deal with out of vocabulary words:

- *sigs – sum of n-grams*
- *sigs- – null vector*

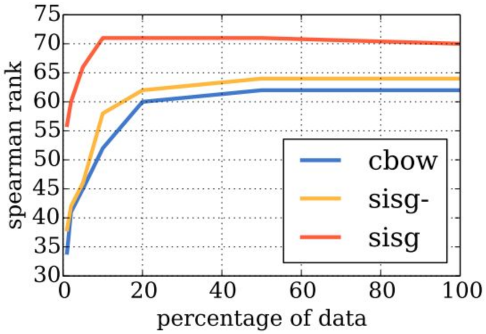
Results. Word analogy tasks

Word analogy task is questions, of the form A is to B as C is to D, where D must be predicted by the models.

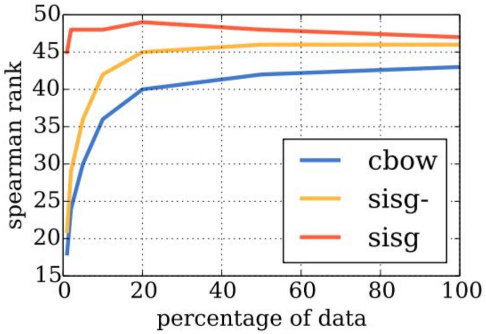
They observe that morphological information significantly improves the syntactic tasks; the approach outperforms the baselines. In contrast, it does not help for semantic questions, and even degrades the performance for German and Italian.

		sg	cbow	sisg
Cs	Semantic	25.7	27.6	27.5
	Syntactic	52.8	55.0	77.8
De	Semantic	66.5	66.8	62.3
	Syntactic	44.5	45.0	56.4
En	Semantic	78.5	78.2	77.8
	Syntactic	70.1	69.9	74.9
It	Semantic	52.3	54.7	52.3
	Syntactic	51.5	51.8	62.7

Results. Effect of the size of the training data



(a) DE-GUR350



(b) EN-RW

The approach is more robust because it is able to model infrequent words: authors used different proportion of the training dataset and measured the performance on the test dataset.

They noticed that:

- For all datasets, and all sizes, the proposed approach (sisg) performs better than the baseline.
- Proposed approach provides very good word vectors even when using very small training datasets.

Results. Summary

- ***Morphologically Rich Languages***: The model works really well with languages that have lots of different word forms, outperforming others because it's great at understanding their complexity.
- ***Handling of Rare Words***: Leveraging n-grams, the model effectively captures information from rare words, ensuring robust and meaningful embeddings..
- ***Training on Limited Data***: Demonstrating efficiency, the model achieves impressive results even with a small proportion of training data, making it suitable for tasks with restricted dataset sizes.

Thank you for your attention!