Investigating Language Universal and Specific Properties in Word Embeddings

---

Authors: Peng Qian, Xipeng Liu & Xuanjing Huang
Speaker: Yupei Du

Outline

- Embedding Models
- Motivations
- Experiment Design
- Results
- Further Analysis

Embedding Models

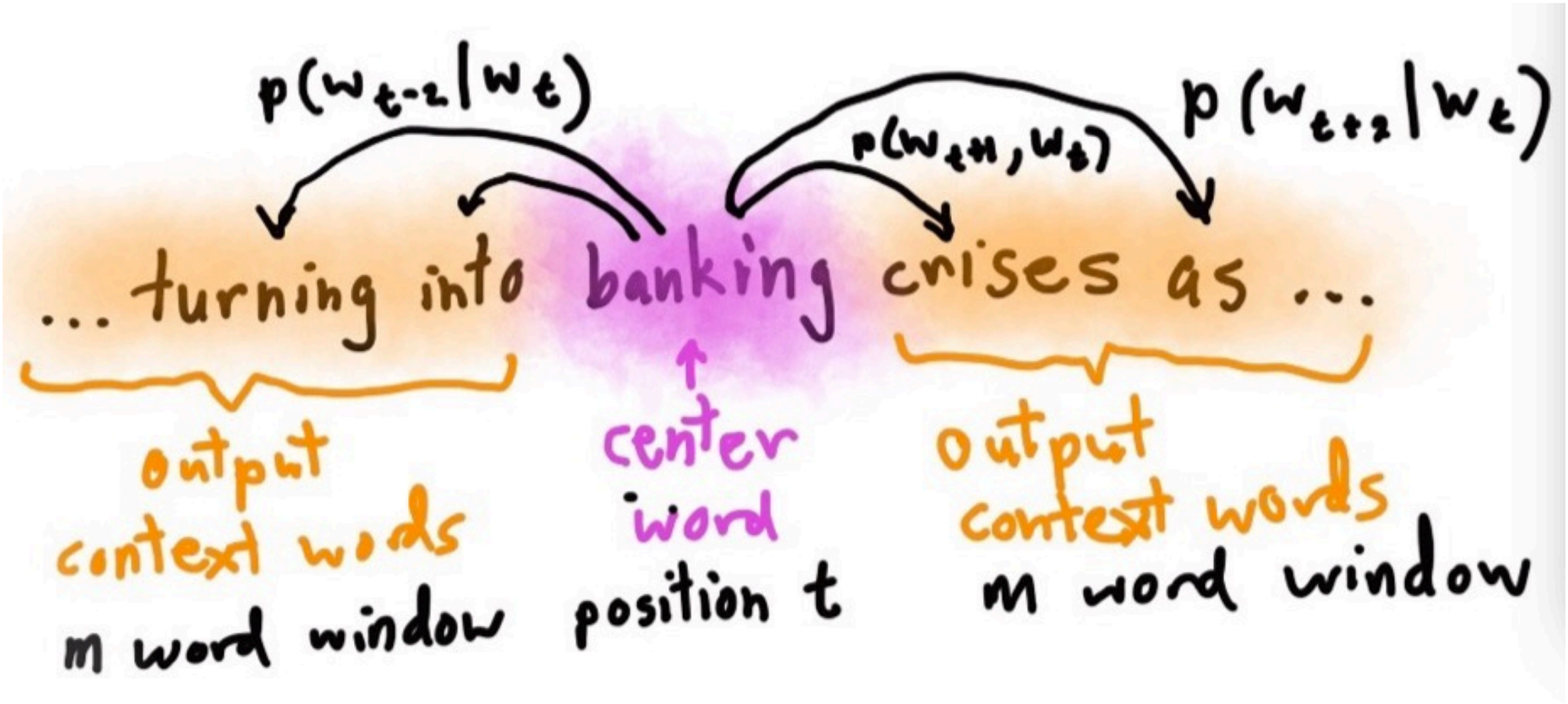
- Word Embedding: represent words in vectors
- C&W Model: keep the context words and their word order information(CW)
- Skip-gram Model: keep the context words information but ignore the word order information(SG)
- Character-Based Model: keep word form information(AE)

C&W Model

- Compute score of sequence $f(w_{t-m}...w_t...w_{t+m})$
- Substitute the center word w_t with different words w_i
- Max the difference between scores
- Loss function: $J(\theta) = \sum_{x \epsilon X} \sum_{w \epsilon V} max\{0, 1 - f_\theta(x) + f_\theta(x^{(w)})\}$

Skip-gram Model

- Predict context words given target(position independent)



Skip-gram Model

- Objective function: $J^{'}(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m}^{j \neq 0} p(w_{t+j}|w_t;\theta)$

- Which equals to: $J(\theta) = \sum_{t=1}^{T} \sum_{-m \leq j \leq m}^{j \neq 0} log(p(w_{t+j}|w_t;\theta))$

- Compute probability: $p(o|c) = \frac{e^{u_0^T v_c}}{\sum_{w=1}^{v} e^{u_w^T v_c}}$

Character-based Models

- Character-based LSTM autoencoder: Takes the character sequence of word as input and reconstruct the input sequence

- Take the hidden layer as word embeddings

- Take the advantage of pure word form instead of the context

Outline

- Embedding Models
- Motivations
- Experiment Design
- Results
- Further Analysis

Motivations

- whether embedding models are immune to the diversity of languages

- Former work:
  empirically interpreting word embedding and exploring the intrinsic and extrinsic factors in learning process

- Understanding model behaviors towards language typological diversity and the utility of context and form for different languages

Topics

- How do typological differences of language structure influence a word embedding model?

- Is word form a more efficient predictor of a certain grammatical function than word context for specific languages?

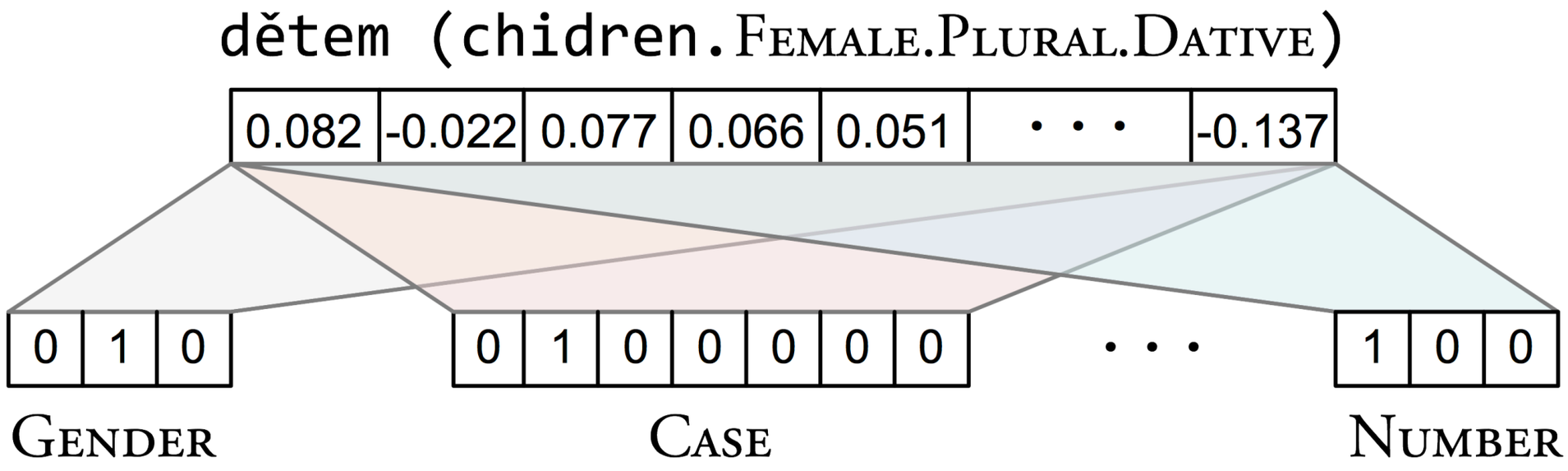
- How do the neurons of a model respond to linguistic features?

Outline

- Embedding Models
- Motivations
- Experiment Design
- Results
- Further Analysis

Experiments Design

- Hypothesis: If word embedding implicitly encode sufficient information, there exist a linear/non-linear map between a word representation x and a high-level sparse feature vector y
- Example: dětem(Czech)



Experiments Design

- 4 series of experiments to compare different models on different languages

ID	Category	Attribute
1	Syntax	Part-of-Speech
2	Syntax	Dependency Relation
3	Morphology	Gender/Number/Case/Animacy/Definite/Person/Tense/ Aspect/Mood/Voice/Pronoun Type/verb form
4	Semantics	Sentiment Score

Experiments Design

- Maps:
  Linear: least L2 error
  Non-Linear: 4 hidden layers MLP(50*80*80*50)

- Syntactic & Morphological
  Normalized label frequency distribution of linguistic attribute a for the word w

- Sentiment feature
  manual collected data transformed to [0,1]

Source

- Universal Dependencies (Version 1.2) and Chinese Treebank (CTB 7.0) for Syntax and Morphology tasks
- Manually annotated data collected by Dodds et al. (2015) native speakers' rates on feelings towards 10000 frequent words

Evaluation

- Emotion Score:
  - Regress the representation to emotion score.
  - Spearman correlations

- Other experiments:
  - predicted tag equals & most probable correct one

$$acc = \frac{1}{|W|}\sum_{i}^{|W|}\Delta(\hat{y}^a_{W_i}, y^a_{W_i}) \qquad (1)$$

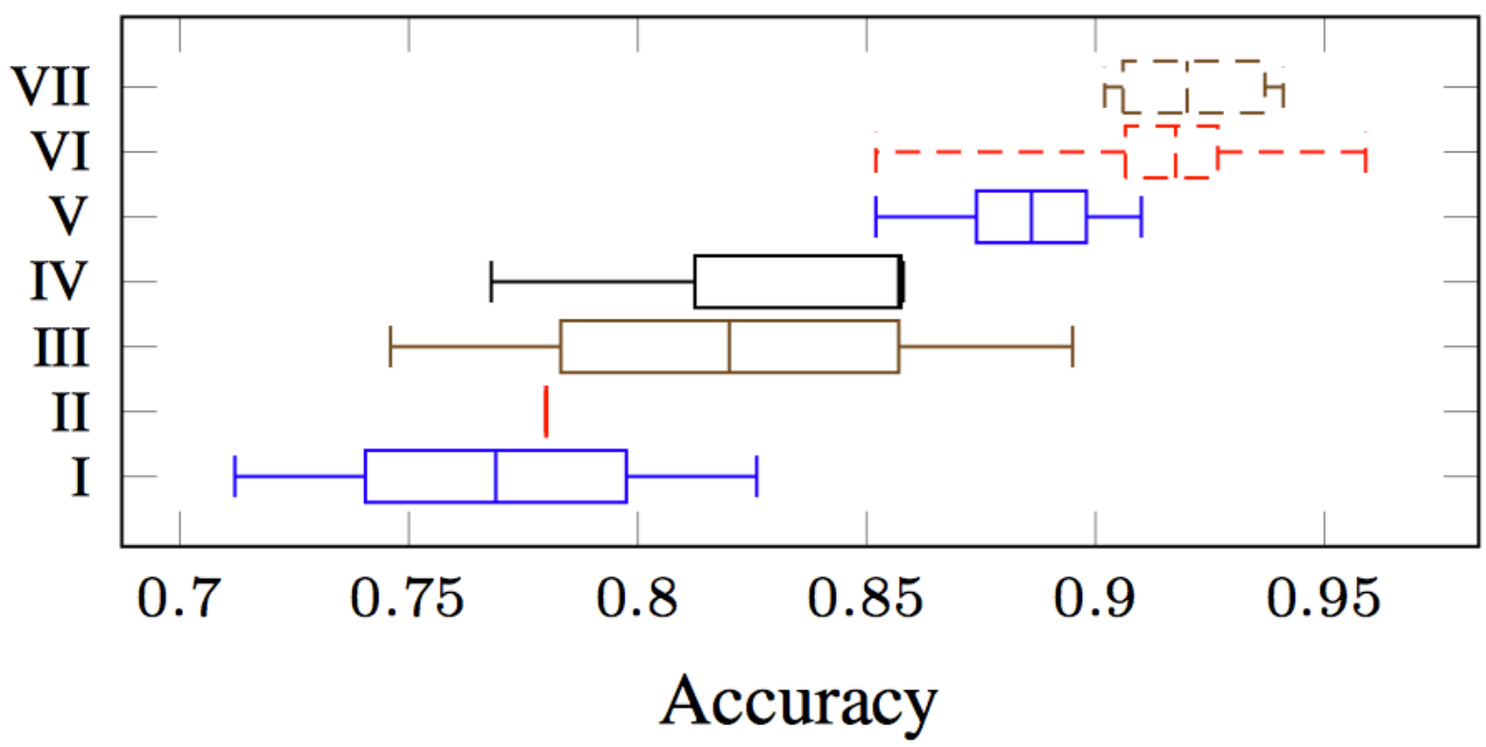
$$\Delta(\hat{y}^a_{W_i}, y^a_{W_i}) = \begin{cases} 1 & \hat{y}^a_{W_i} = y^a_{W_i} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Outline

- Embedding Models
- Motivations
- Experiment Design
- Results
- Further Analysis

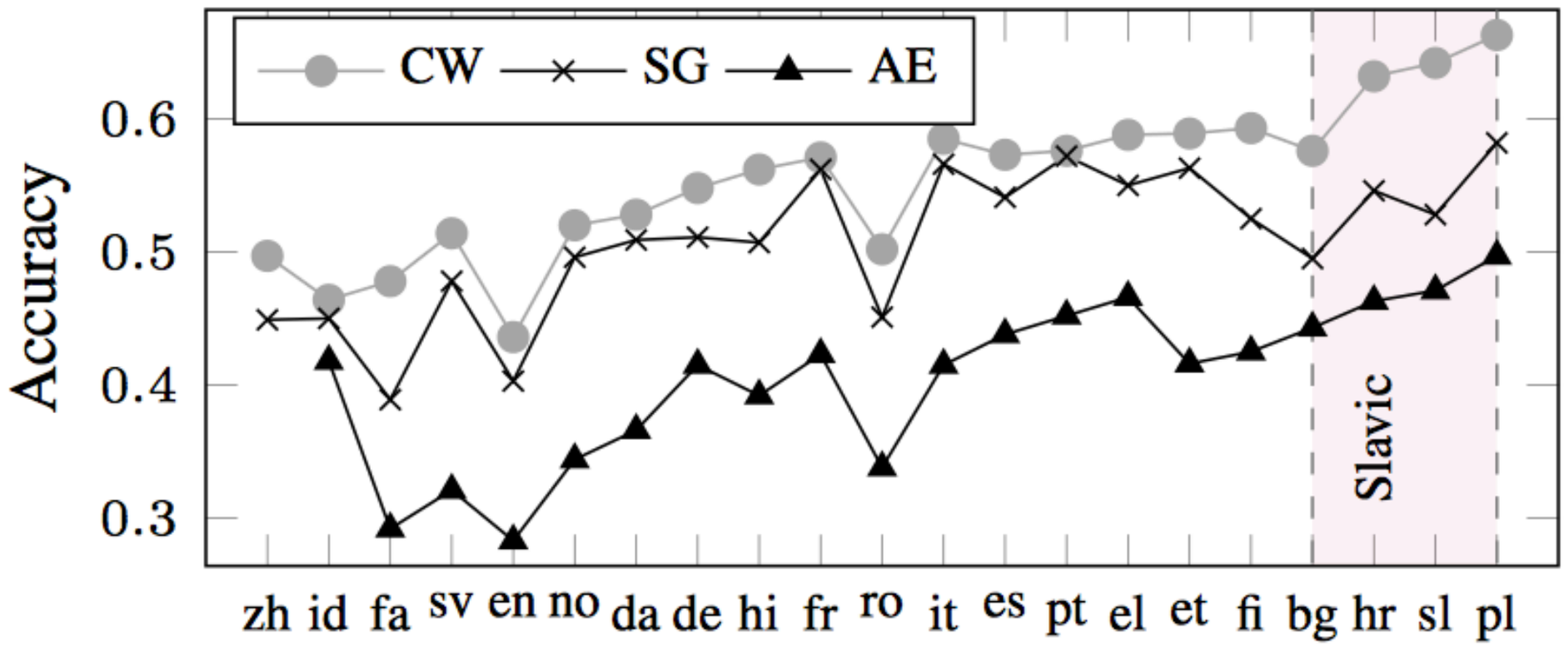
Part-of-Speech

- Context-based embeddings performs better (CW > SG)
- AE performs good on languages employ affix makers to indicate POS category(Russian, Czech and Indonesian)
- CW perform similar on languages of same type of word orders



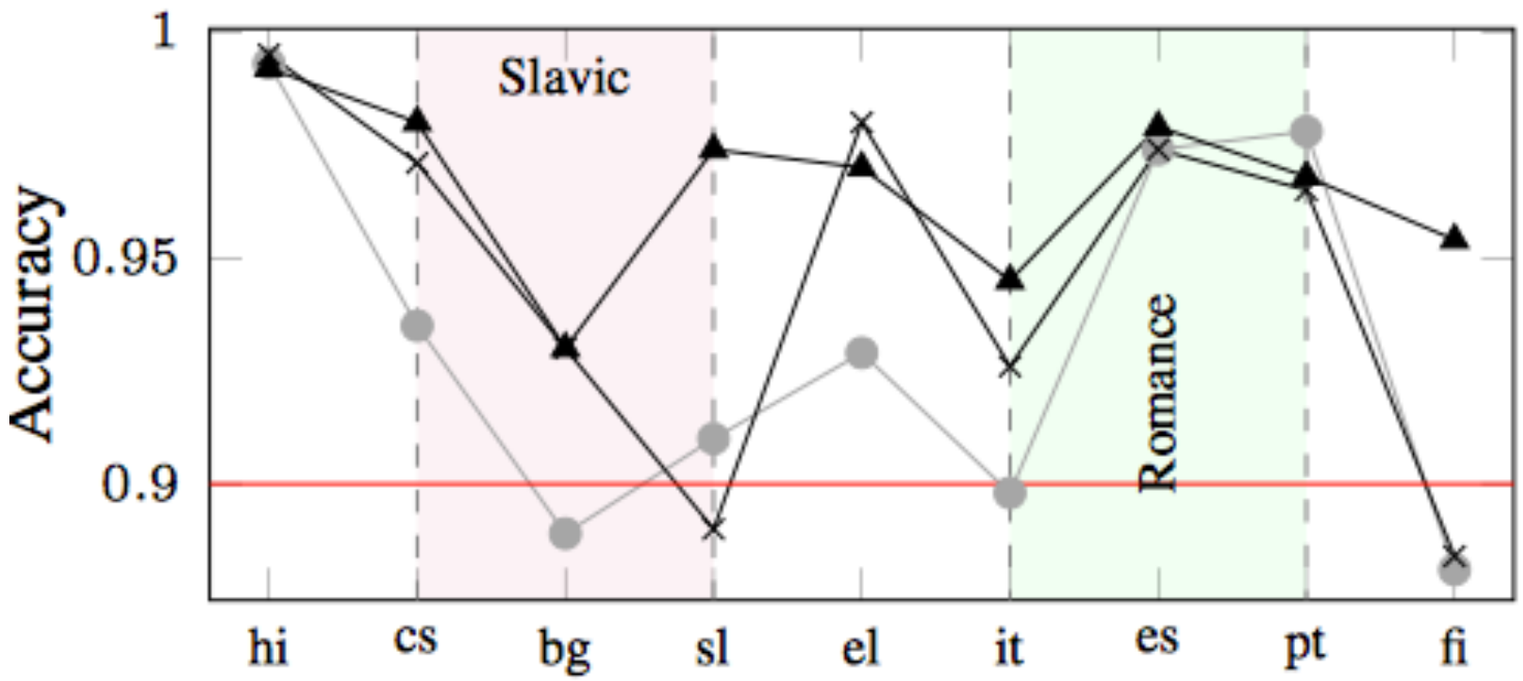
Dependency Relation

- Overall performance is worser than POS
- CW achieve best performance



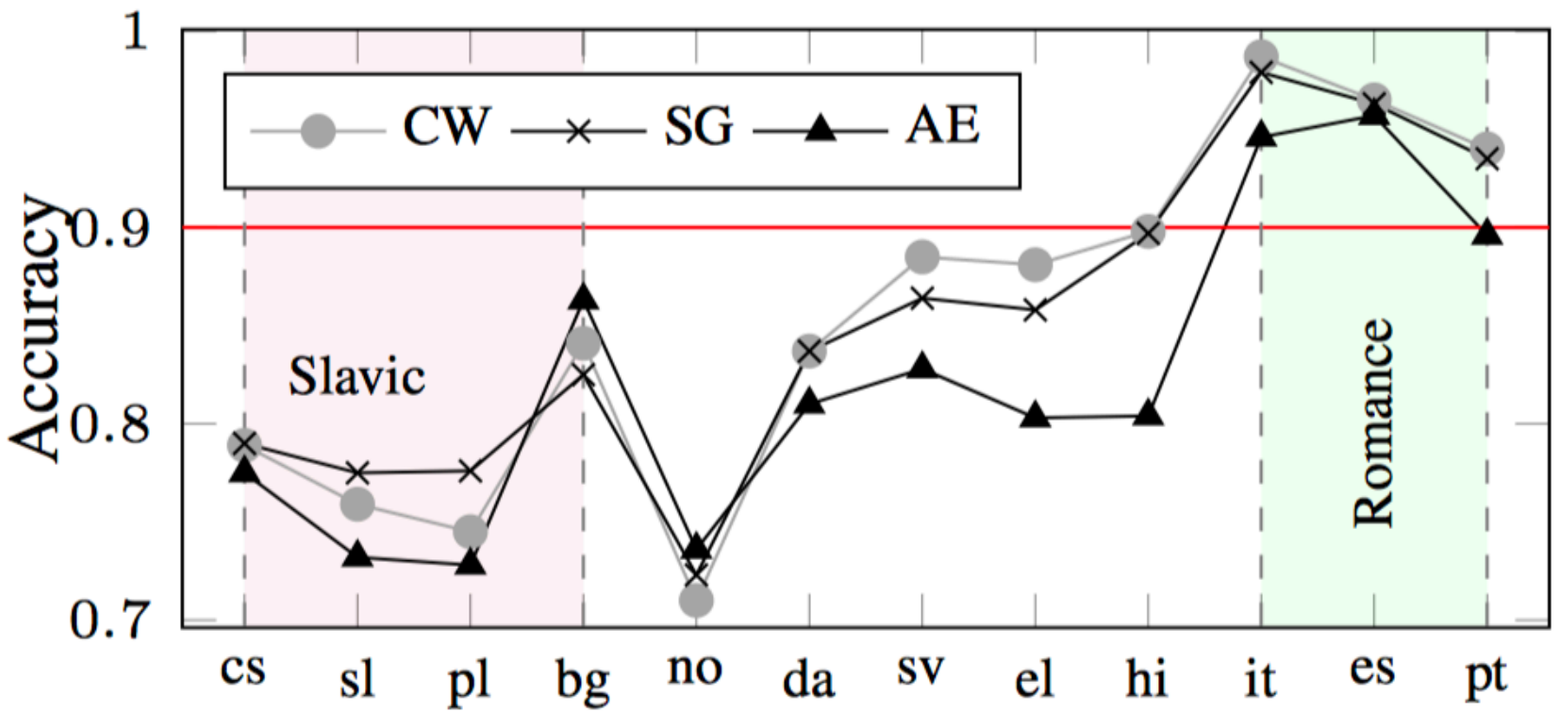
Morphological Features

- Split into 2 kinds, 5 nominal features (gender, number, case, animacy, definiteness) and 7 verbal features (person, tense, aspect, voice, pronoun type, verb form)
- All 3 models give perfect performance on verbal features



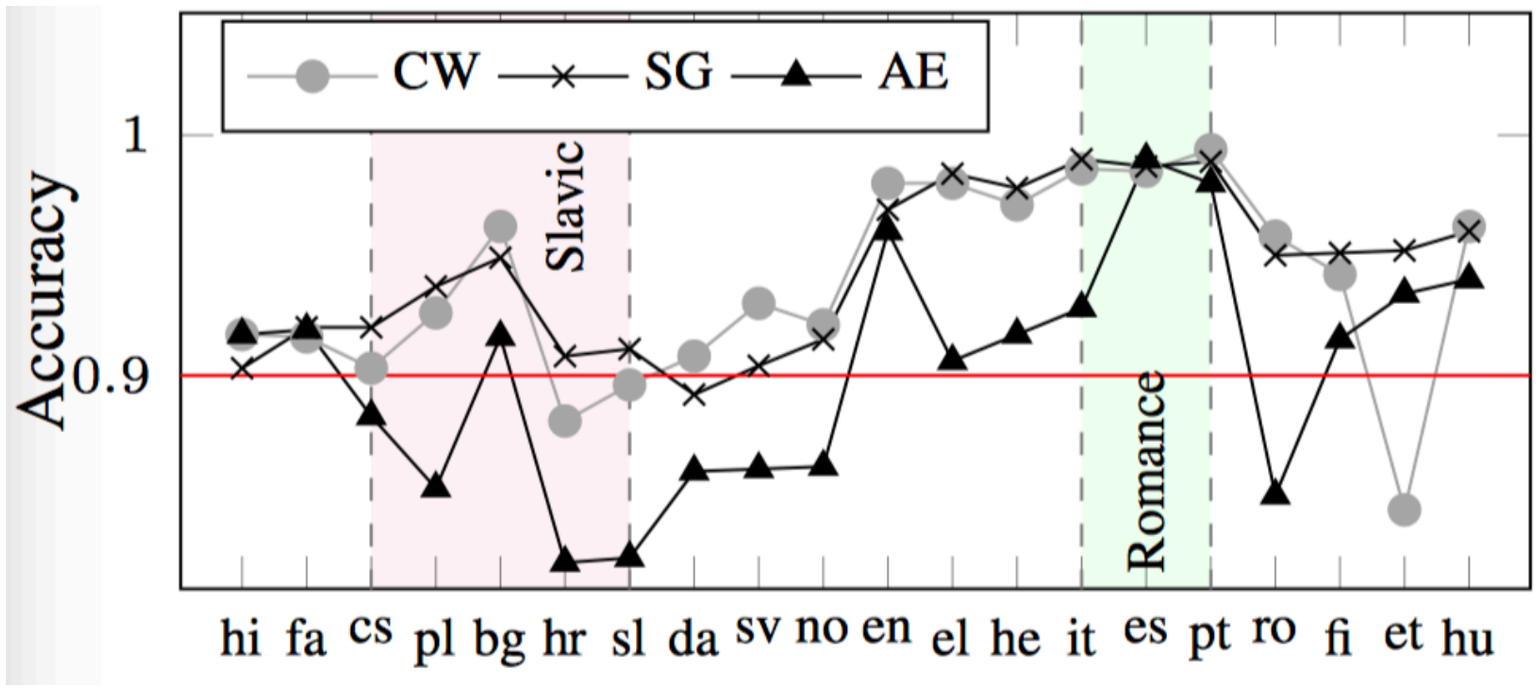
Gender

- Partially based on semantics and agreements between noun and determiners but also expressed via declension and umlaut
- Romance language employ regular rules to judge, Slavic language have nonlinear feature hard to tackle



Number

- Linguistic abstraction of objects' quantities
- Inflectional feature of parts of speech
- All 3 perform well, AE perform as good as CW and SG on English, Spanish and Portuguese



Case

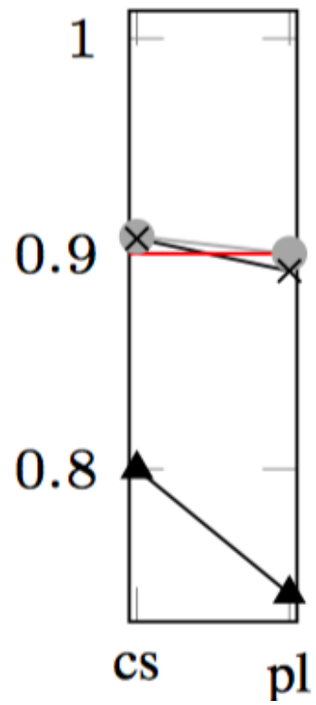
	Language	$\|V\|$	C&W	SG	AE	# Case
Analy.	Danish	372	0.947	0.946	1.000	3
	Swedish	5893	0.995	0.990	0.981	2
	Bulgarian	104	0.636	0.546	0.818	4
Agglut.	Finnish	21094	0.868	0.871	**0.908**	15
	Hungarian	4536	0.852		**0.901**	22
	Tamil	1144	0.896	–	**0.835**	7
	Basque	8020	0.761	–	**0.857**	15
Fusional	Hindi	10682	0.712	0.704	0.646	7
	Czech	38666	0.788	0.776	0.663	7
	Polish	13715	0.828	0.785	0.636	7
	Slovenian	15150	0.796	0.768	0.617	6
	Croatian	9945	0.807	0.789	0.628	7
	Greek	5790	0.841	0.851	0.774	5
	Latin	4773	0.674	–	0.636	7

Case

- Relational morpheme explicitly reflects semantic role of noun with inflection
- Swedish have only 2 cases: nominal and genitive
- in AE, fusional languages perform worse than agglutinative languages
- Case system is simplified in analytic languages
- Cases are special semantic relations around target

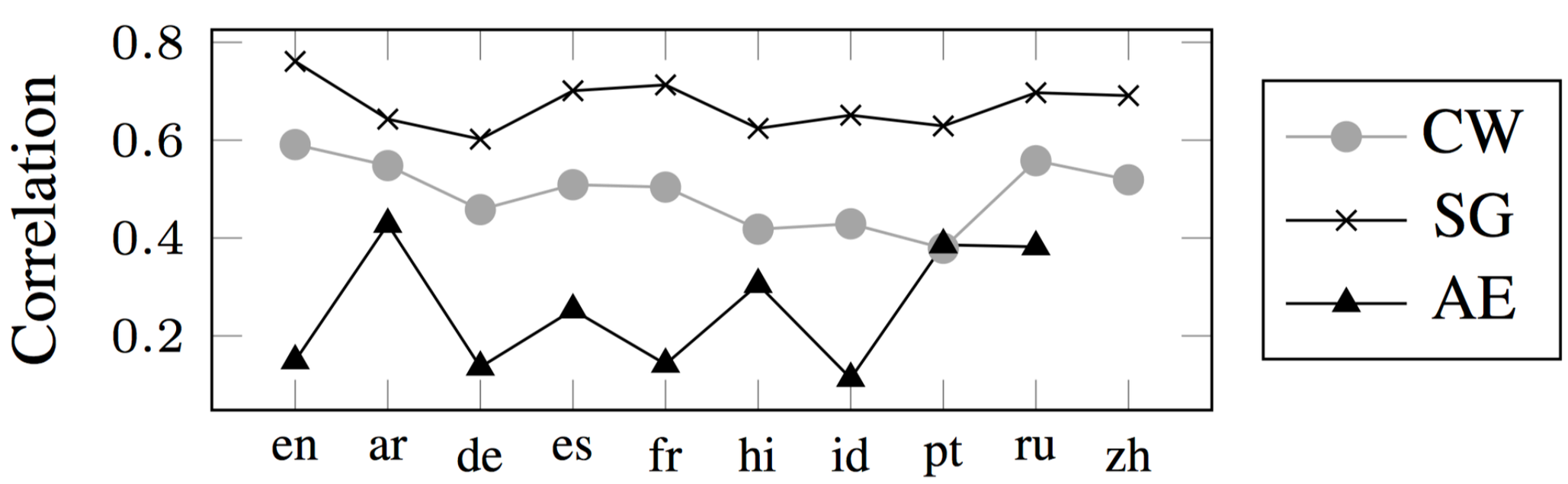
Animacy

- Determine animate objects from inanimate ones.
- Lexical semantic feature



Emotion Score

- Emotions usually come from the context and can hardly get from the word form

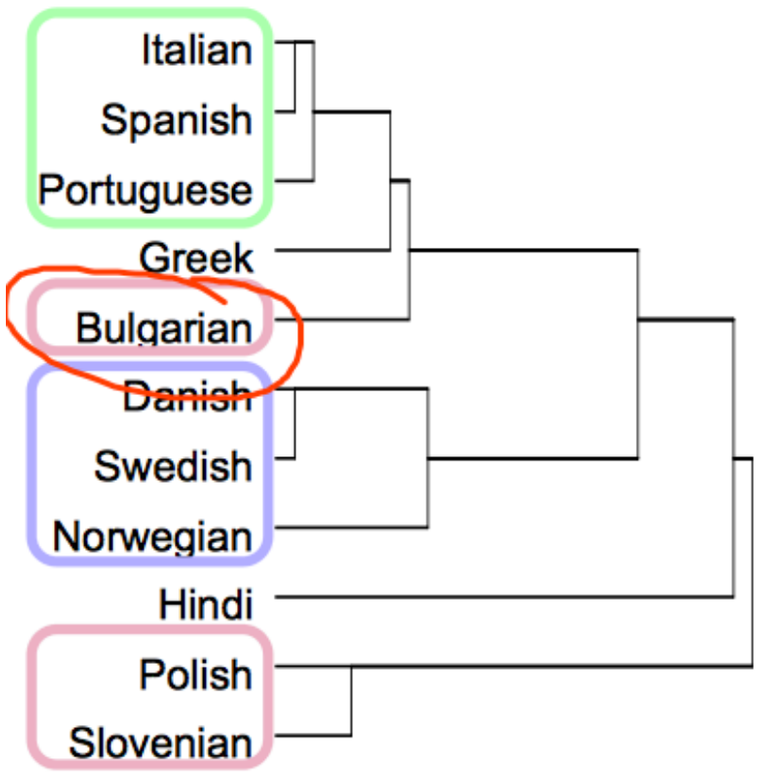


Outline

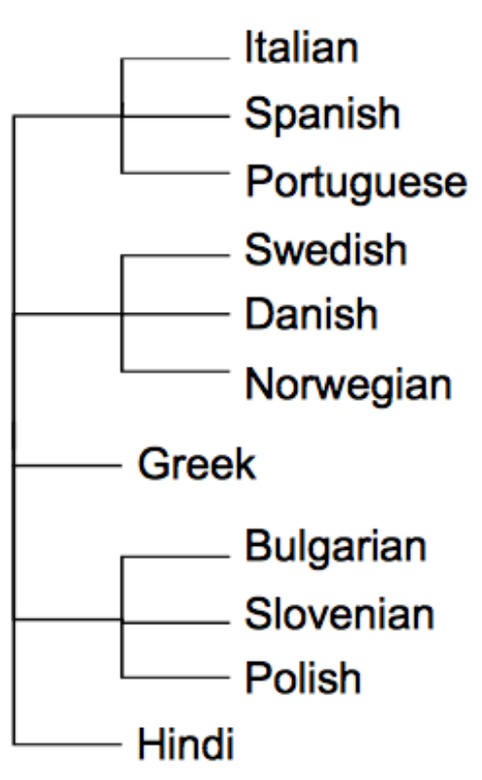
- Embedding Models
- Motivations
- Experiment Design
- Results
- Further Analysis

Typology vs. Phylogeny

- Cluster languages with model performance variation



(a) Hierarchical tree based on model performances



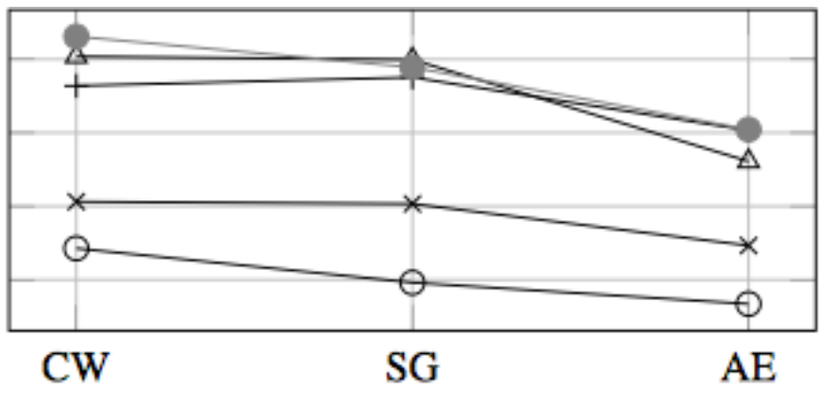
(b) WALS Genus Tree

Form vs. Context

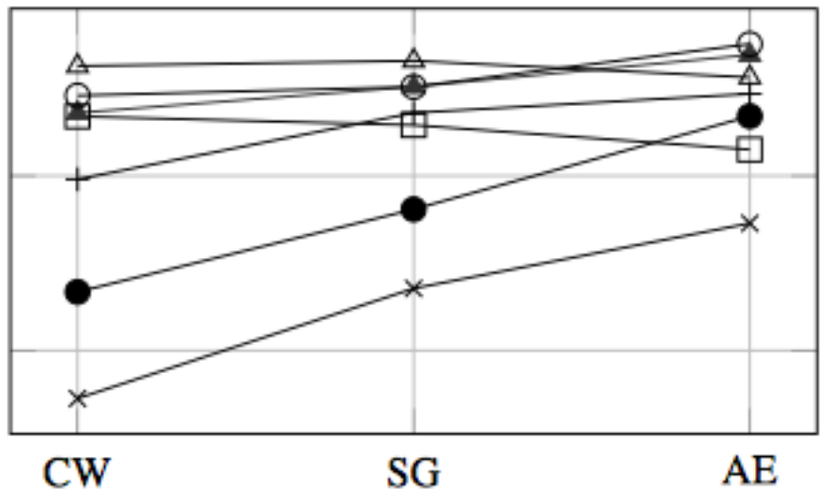
- Word order is vital to syntactic information but kind of noise to semantic information

- Word form could be cue to decode morphological features

Nominal



Verbal



Form vs. Context

- Autoencoder on shuffled character-based word representation
- Word form is necessary for learning grammatical function

POS tag

Lan	Raw	Shuffled
Russian	0.906	0.671
Slovenian	0.800	0.653

Form vs. Context

- Many languages share similar phonographic writing system

- Finnish-English / Arabic-Persian(Urdu)

Source Language	**Arabic**		**Finnish**		
Target Language	fa	ud	en	shuf en	rand
Bigram type overlap.	0.176	0.761	0.891	0.864	0.648
Bigram token overlap.	0.689	0.881	0.999	0.993	0.650
Trigram type overlap.	0.523	0.522	0.665	0.449	0.078
Trigram token overlap.	0.526	0.585	0.978	0.796	0.078
Reconstruction Acc.	0.586	0.689	0.95	0.83	0.22

Form vs. Context

- Phonological structures in writing system
- Character-based Model memorize the grapheme or phoneme clusters of words
- Models trained on Finnish increase accuracy of POS tag task.

Neuronal Activation Pattern

- Grandmother neuron
- Selectivity P

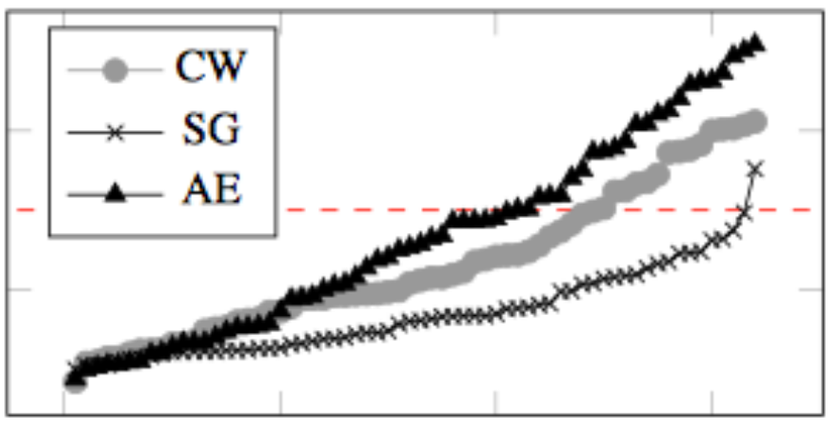
$$c_{f,t} = \frac{N^{+}_{f,t}}{N_f}, c_{\neg f,t} = \frac{N^{+}_{\neg f,t}}{N_{\neg f}},$$

$$Selectivity = p_{f,t} = \frac{2 \times c_{f,t} \times c_{\neg f,t}}{c_{f,t} + c_{\neg f,t}}$$

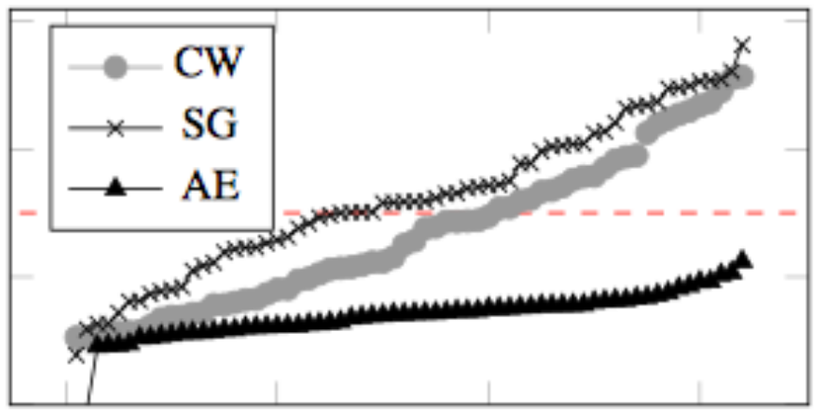
Neuronal Activation Pattern

- Character-based model capture more morphological information
- Word-based model capture both semantic and morphological information

me—kan



Country names



Thank you!