Multilingual Grammar Induction with Continuous Language Identification

Wenjuan Han, Ge Wang, Yong Jiang, Kewei Tu

ShanghaiTech University, Shanghai, China
Alibaba Group
{hanwj,wangge,tukw}@shanghaitech.edu.cn
{yongjiang.jy}@alibaba-inc.com

November 9, 2019

Outline

Outline

- Grammar induction is the task to learn grammars form unannotated corpus.

- Grammar induction is the task to learn grammars form unannotated corpus.
- Multilingual grammar induction couples grammar parameters of different languages together and learns them simultaneously.

- Grammar induction is the task to learn grammars form unannotated corpus.
- Multilingual grammar induction couples grammar parameters of different languages together and learns them simultaneously.

  $\rightarrow$ The key is to exploit the similarities between languages.

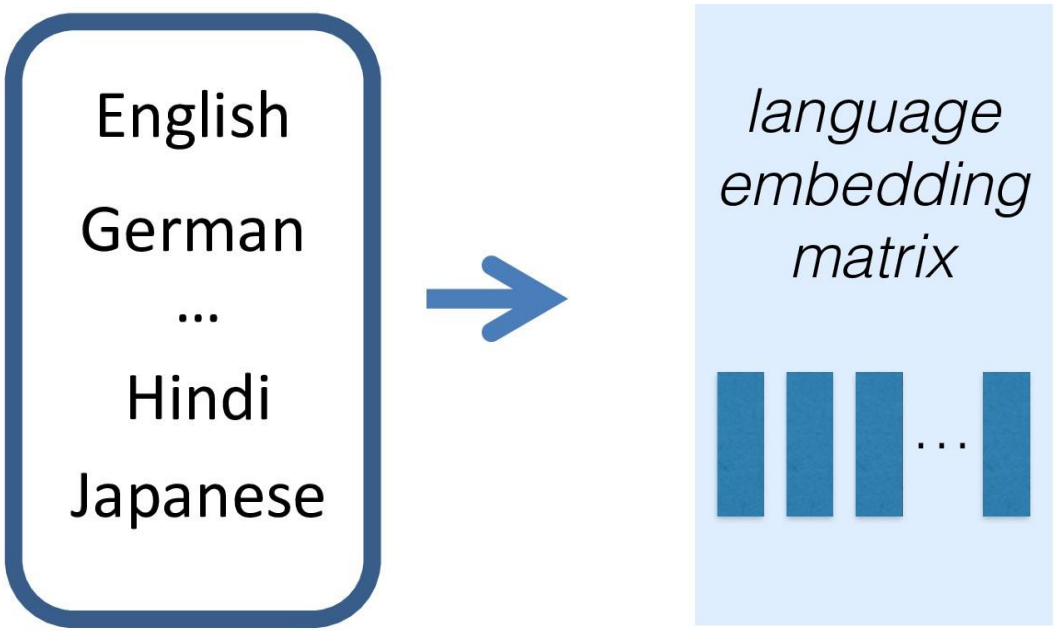
Existing approaches to tackle this problem:

- Treating languages equally (Iwata et al., 2010).
- Utilizing hand-crafted phylogenetic tree to encode this kind of information (Berg-Kirkpatrick and Klein, 2010).

Existing approaches to tackle this problem:

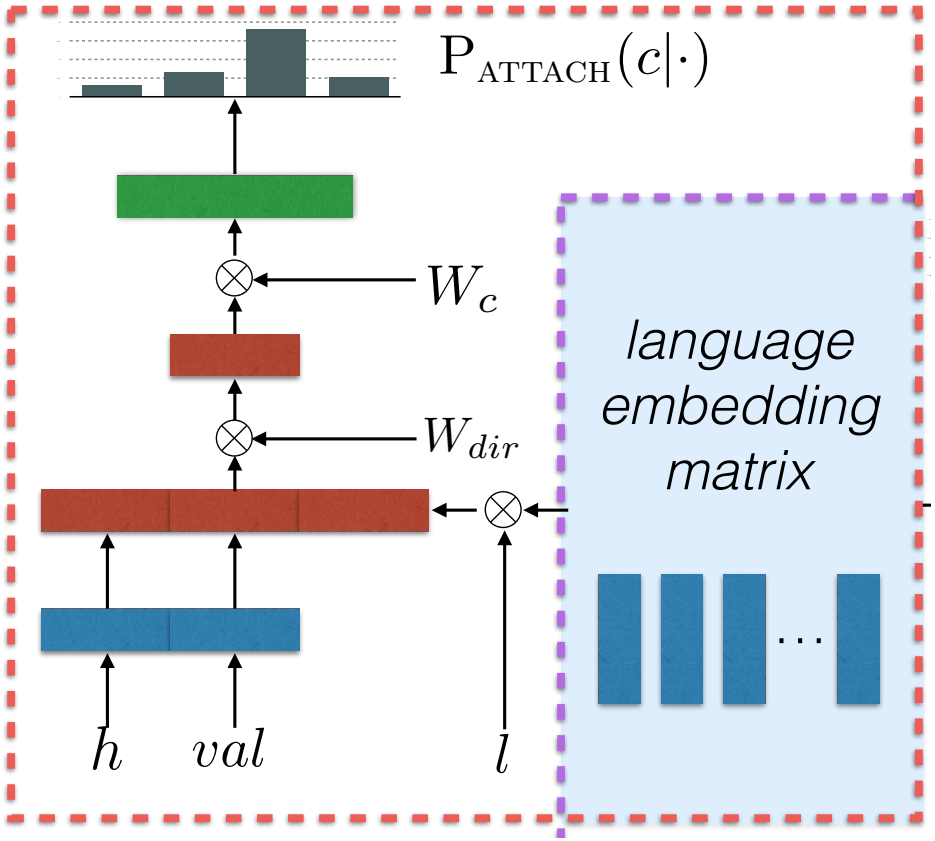
- Treating languages equally (Iwata et al., 2010). → Language similarity ignored.
- Utilizing hand-crafted phylogenetic tree to encode this kind of information (Berg-Kirkpatrick and Klein, 2010). → Need linguistic knowledge and sometimes could be misleading. Example: English is dominant SVO while German is not, although they are both Germanic languages.

Outline



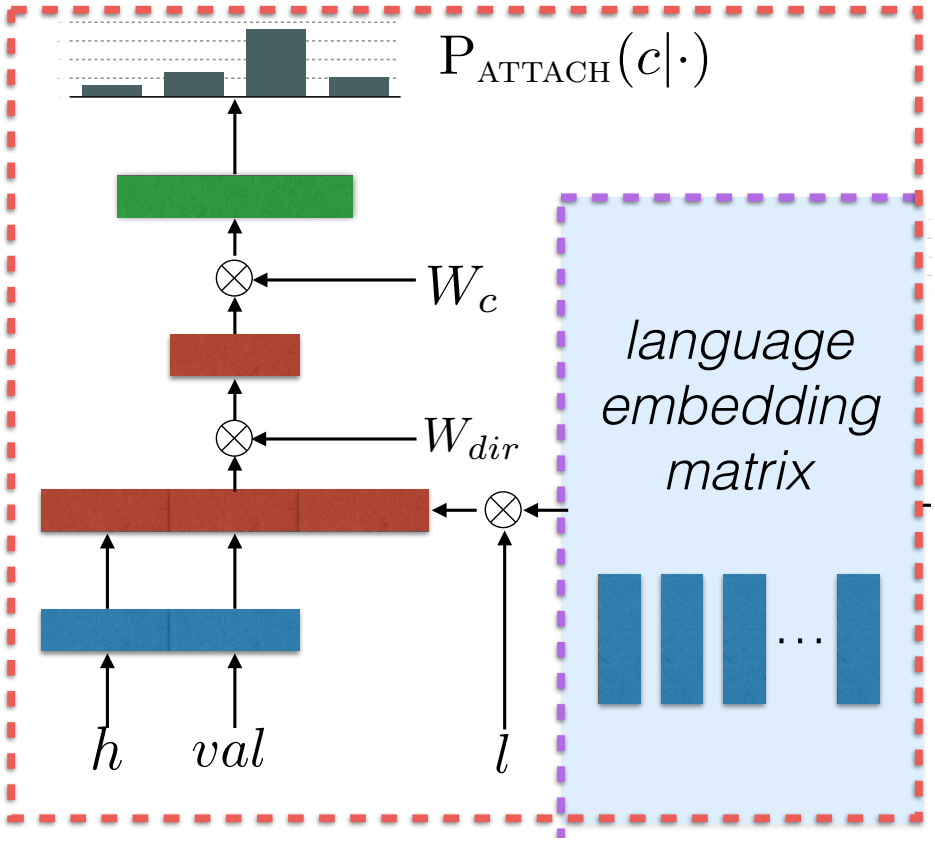
We represent language identities with continuous vectors i.e., language embeddings and use them to encode language similarity.

Model Architecture



Neural DMV grammar rule probability:
$P_{ATTACH}(child|head, direction, valence)$

Model Architecture

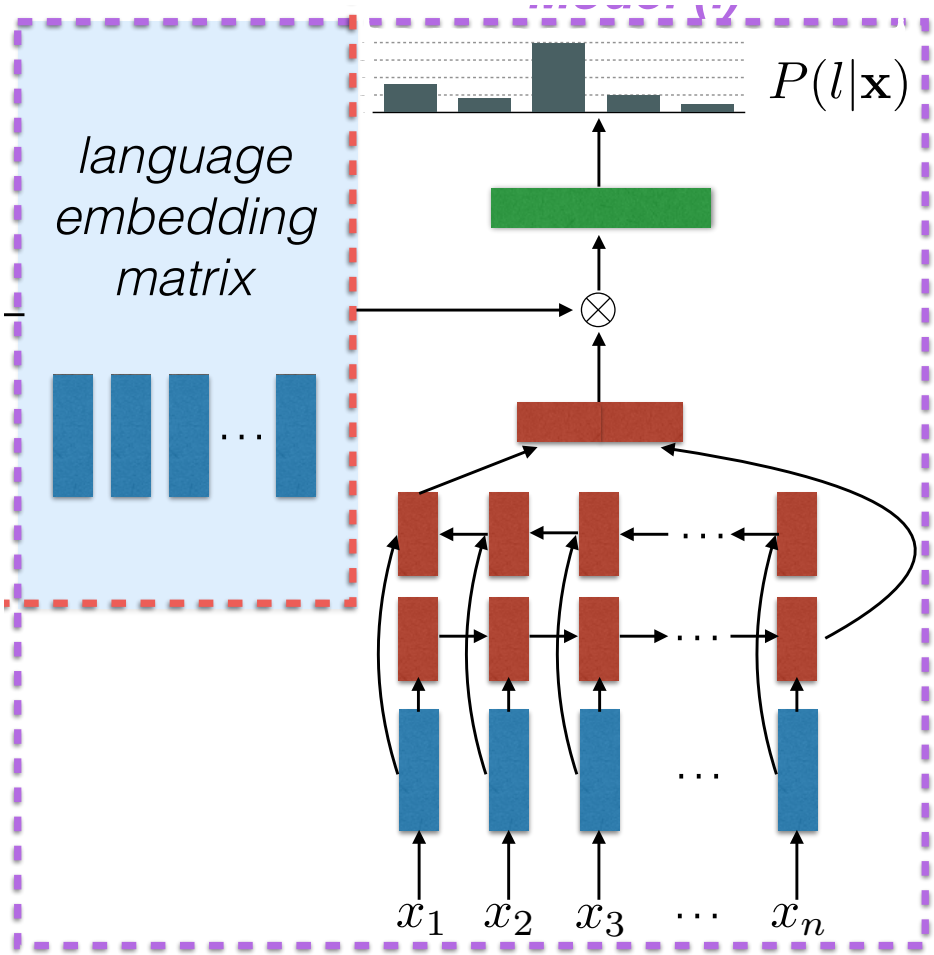


Neural DMV grammar rule probability:

$P_{ATTACH}(child|head, direction, valence)$

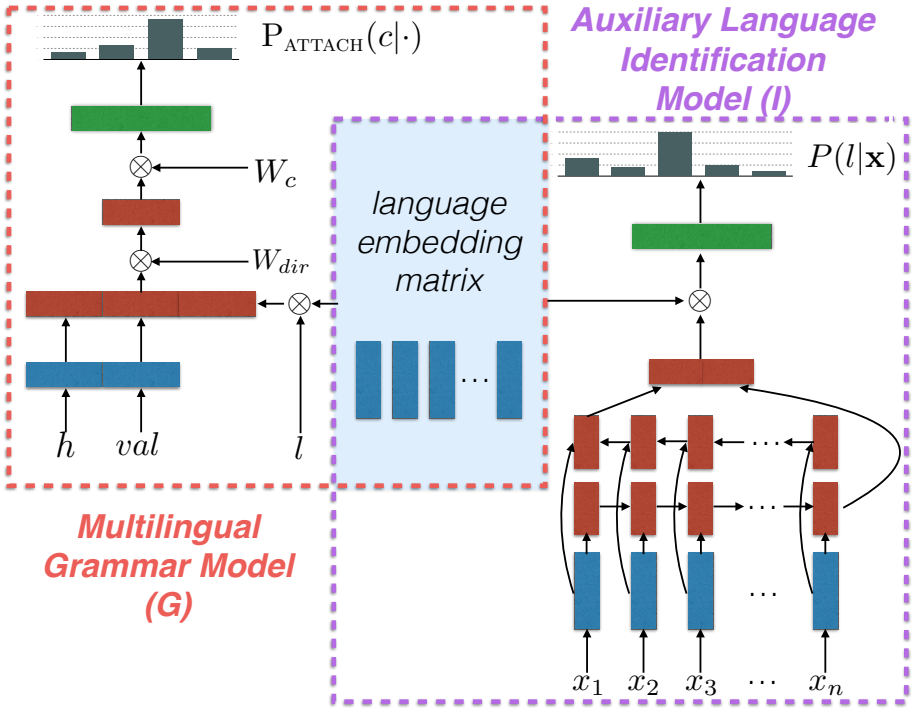
Now we have: $P_{ATTACH}(child|head, direction, valence, language)$

Model Architecture



Predict language identification with language embeddings and sentence representations.

Model Architecture



For each training sentence $\mathbf{x}^{(i)}$ from language l:

- $P(\mathbf{x}^{(i)}|\mathbf{G}_{l^{(i)}})$, the probability of the training sentence $\mathbf{x}^{(i)}$ being generated from grammar $\mathbf{G}_{l^{(i)}}$.
- $P(l^{(i)}|\mathbf{x}^{(i)})$, the probability of correct language identification of $\mathbf{x}^{(i)}$.

Outline

Objective

For each training sentence $\mathbf{x}^{(i)}$:

- $P(\mathbf{x}^{(i)}|\mathbf{G}_{l^{(i)}})$, the probability of the training sentence $\mathbf{x}^{(i)}$ being generated from grammar $\mathbf{G}_{l^{(i)}}$.
- $P(l^{(i)}|\mathbf{x}^{(i)})$, the probability of correct language identification of $\mathbf{x}^{(i)}$.

The training objective is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{(\mathbf{x},l)\in\mathcal{D}} \Big(\log P_{\boldsymbol{\Theta}}(\mathbf{x}|\mathbf{G}_l) + \lambda \log P_{\boldsymbol{\Theta}}(l|\mathbf{x}) \Big)$$

Learning

- $P(\mathbf{x}^{(i)}|\mathbf{G}_{l^{(i)}}) \rightarrow$ this term is optimized with EM (Adam used in M step).
- $P(l^{(i)}|\mathbf{x}^{(i)}) \rightarrow$ Adam to optimize this term.

Outline

Dataset

We selected 15 languages across 8 language families and subfamilies from UD dataset to ensure diversity.

Language	UD Treebank	Language Family	Corpus Size
ET	Estonian	Finnic	11404
FI	Finnish	Finnic	9648
NL	Dutch	Germanic	8783
EN	English	Germanic	7674
DE	German	Germanic	7447
NO	Norwegian	Germanic	10017
GRC	Ancient_Greek	Hellenic	9387
HI	Hindi	Indo-Iranian	4997
JA	Japanese	Japonic	7441
FR	French	Romance	4976
IT	Italian	Romance	6492
LA	Latin-ITTB	Romance	10136
BG	Bulgarian	Slavonic	6507
SL	Slovenian	Slavonic	3800
EU	Basque	Vasconic	4271

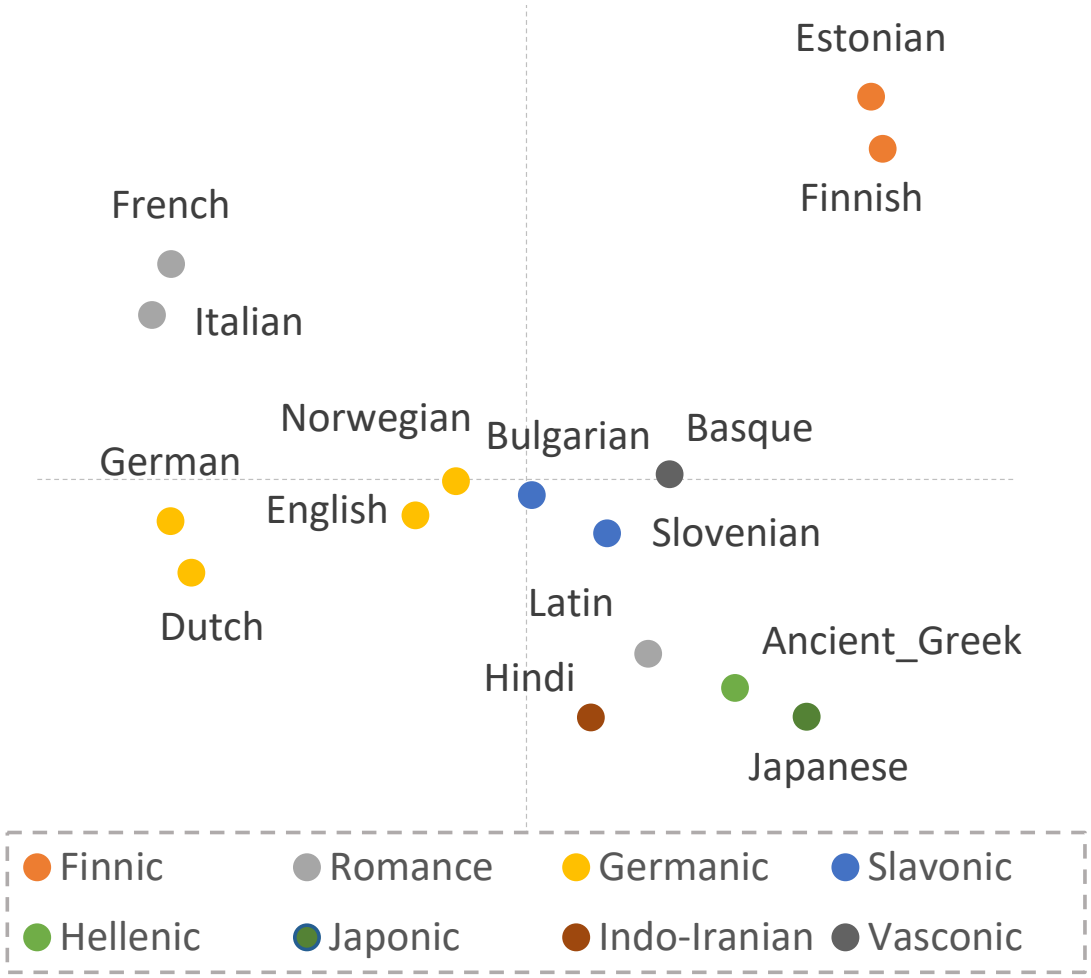
Comparison of monolingual and multilingual approaches.

- G: our multilingual grammar model.
- G+I: our multilingual grammar model and auxiliary language identification task.

Code	Monolingual		Multilingual			
	DMV	**NDMV**	**DMV**	**NDMV**	**G**	**G+I**
ET	51.8	52.9	43.1	45.3	56.0	**56.4**
FI	31.8	27.6	39.1	40.0	**50.7**	49.3
NL	42.4	35.6	46.5	47.8	50.4	**50.6**
EN	51.8	**53.7**	47.7	50.8	51.7	52.7
DE	52.8	50.4	55.5	57.2	59.6	**61.4**
NO	58.9	59.2	55.7	58.8	61.0	**61.3**
GRC	40.4	37.7	41.1	40.8	**46.8**	46.2
HI	52.6	**53.9**	29.2	31.1	47.4	46.8
JA	39.8	37.1	27.8	29.6	43.4	**44.2**
FR	58.8	38.1	59.6	59.4	58.4	**60.1**
IT	60.8	63.6	**66.7**	66.4	64.4	65.9
LA	32.6	36.3	39.8	42.0	**45.1**	45.0
BG	58.9	61.8	65.9	69.4	**71.3**	**71.3**
SL	**70.7**	67.5	62.1	63.3	68.3	68.6
EU	42.1	45.5	45.7	45.2	**54.2**	53.6
Avg	49.7	48.1	48.4	49.8	55.3	**55.6**

Each language is indicated by its ISO 639 code.

Visualization of the language embeddings



Outline

Conclusion

- We represent language identities with language embeddings and use them to encode language similarity.
- The language embeddings are used for grammar parameter prediction and auxiliary language identification task.
- The language embeddings learned in our model can capture language similarity that can not be inferred from phylogenetic knowledge.

Multilingual Grammar Induction with Continuous Language Identification

Wenjuan Han, Ge Wang, Yong Jiang, Kewei Tu

ShanghaiTech University, Shanghai, China
Alibaba Group
{hanwj,wangge,tukw}@shanghaitech.edu.cn
{yongjiang.jy}@alibaba-inc.com

November 9, 2019