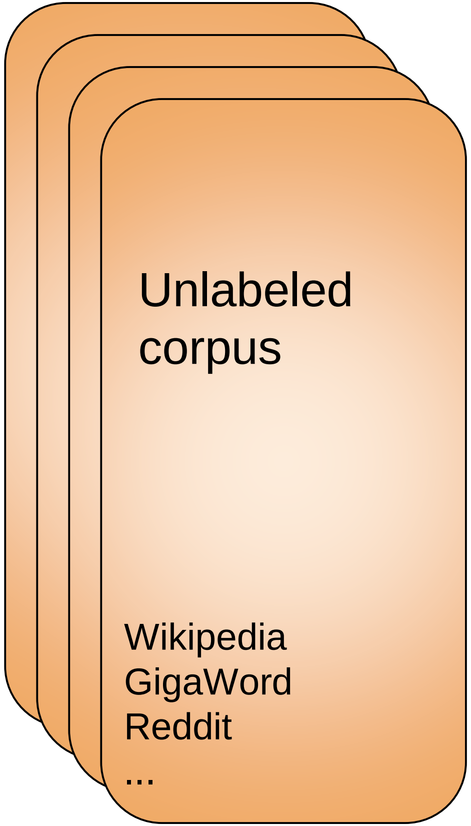
Mimicking Word Embeddings using Subword RNNs

Yuval Pinter, Robert Guthrie, Jacob Eisenstein

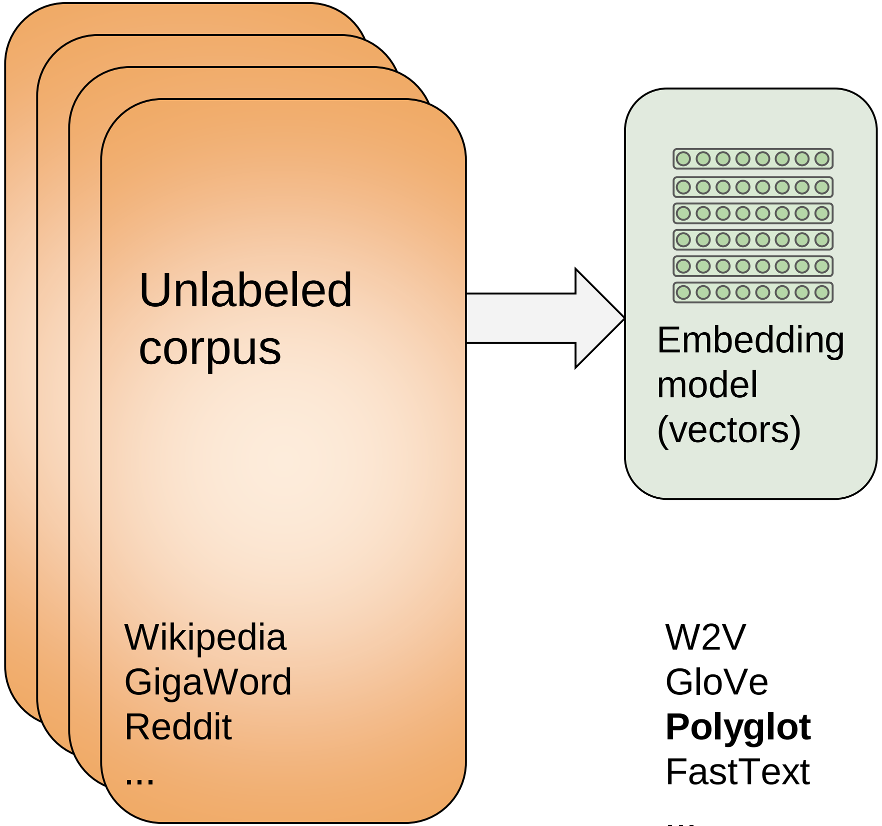
@yuvalpi

Presented at EMNLP
September 2017, Copenhagen

The Word Embedding Pipeline

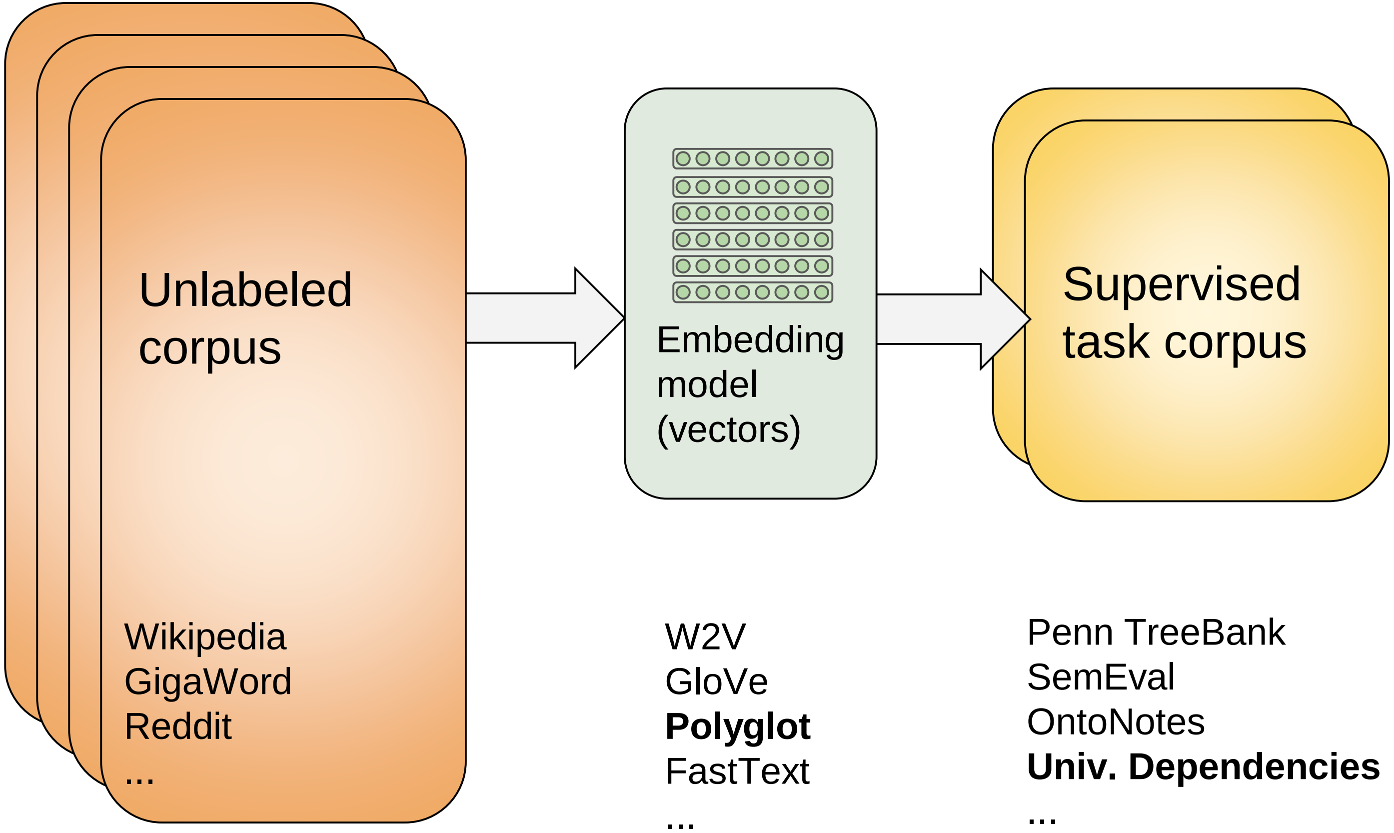


The Word Embedding Pipeline

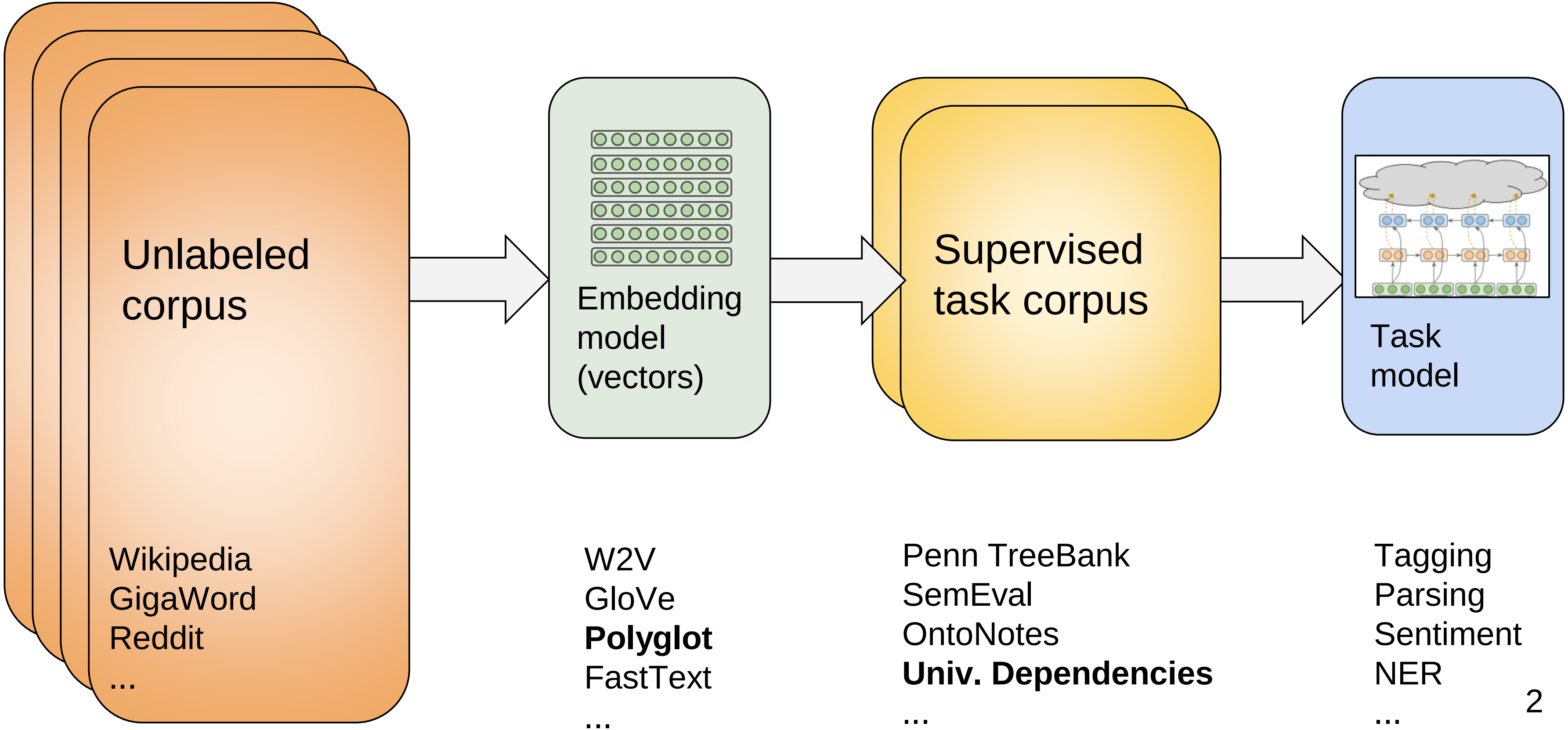


W2V
GloVe
Polyglot
FastText
...

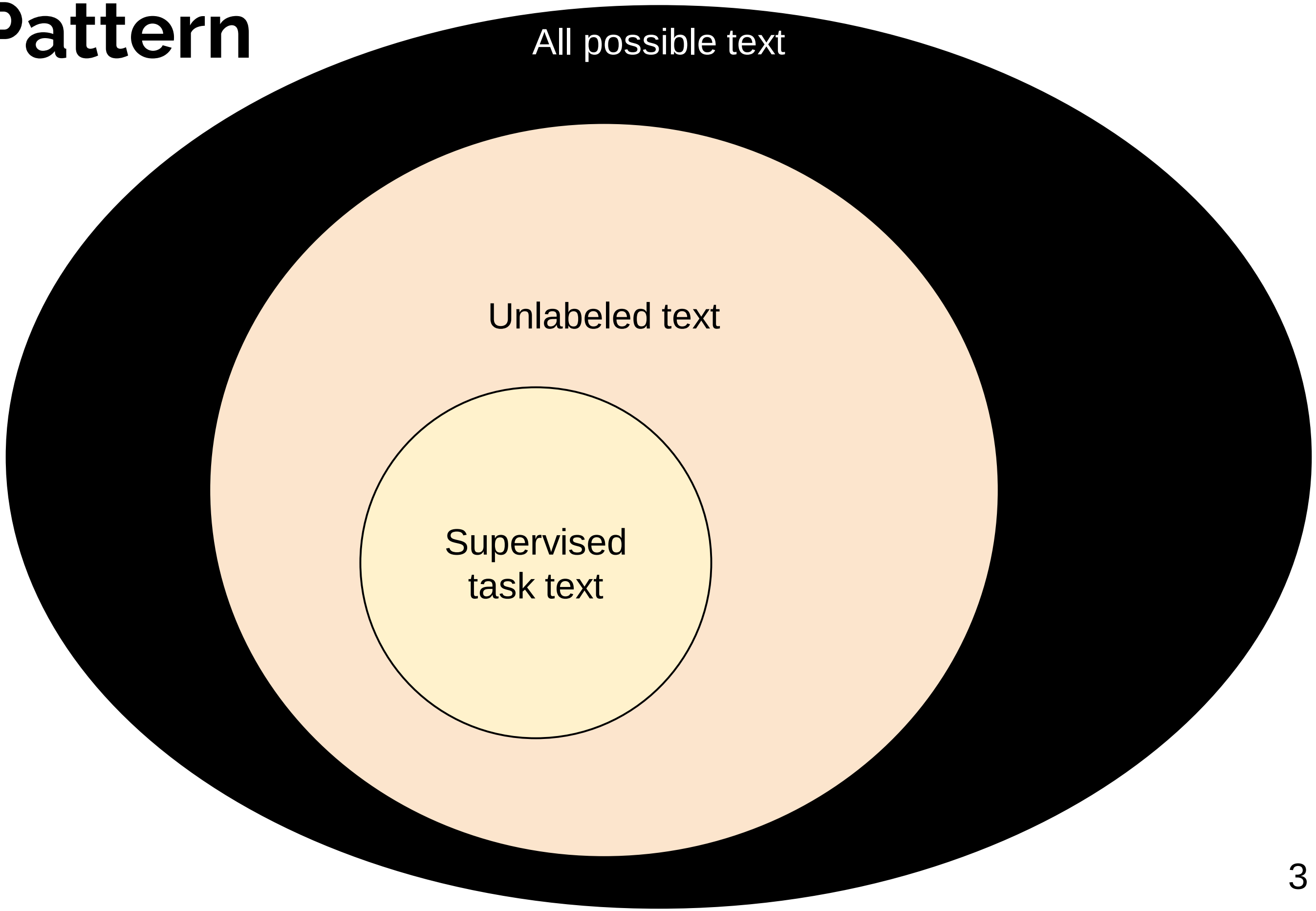
The Word Embedding Pipeline



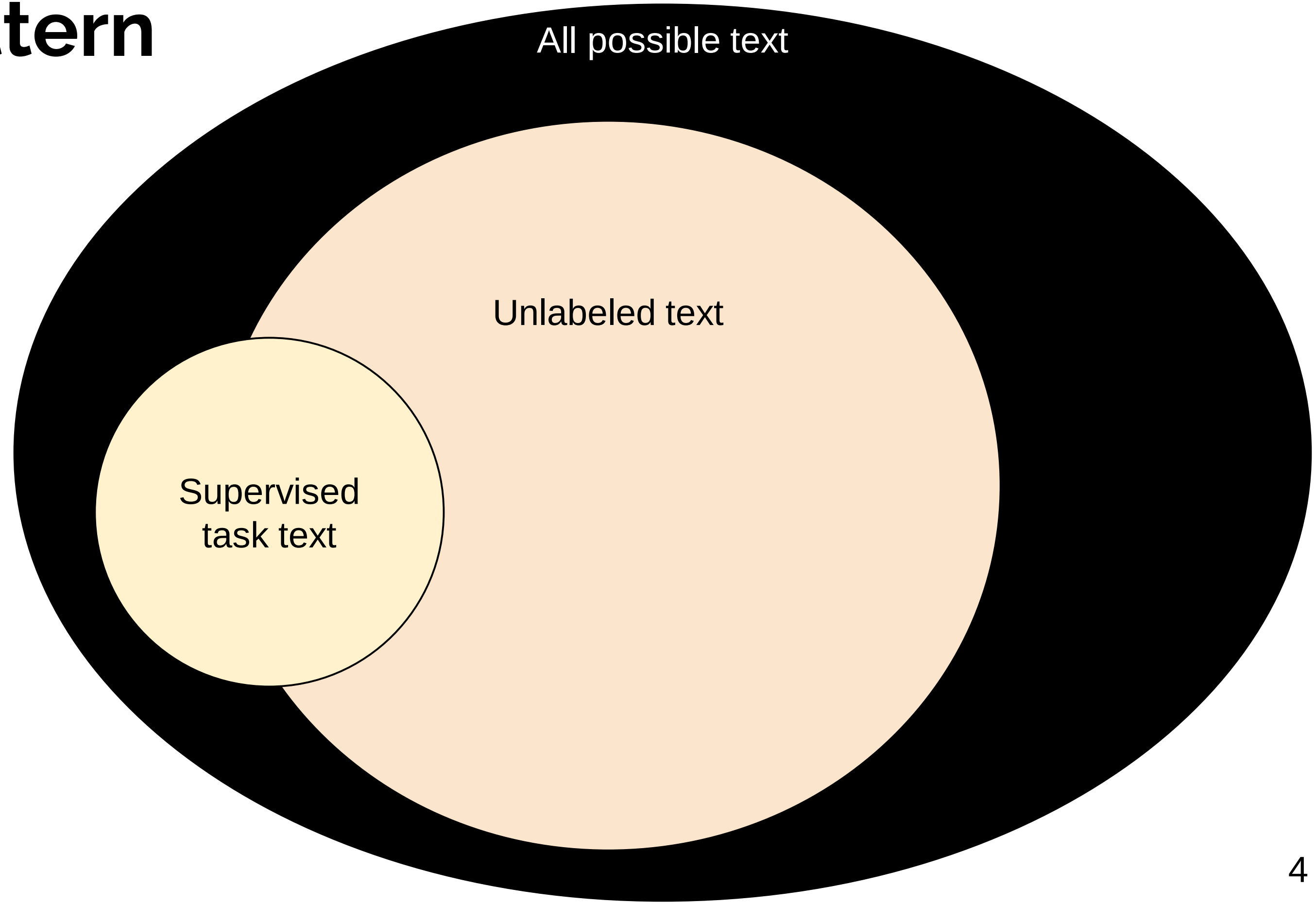
The Word Embedding Pipeline



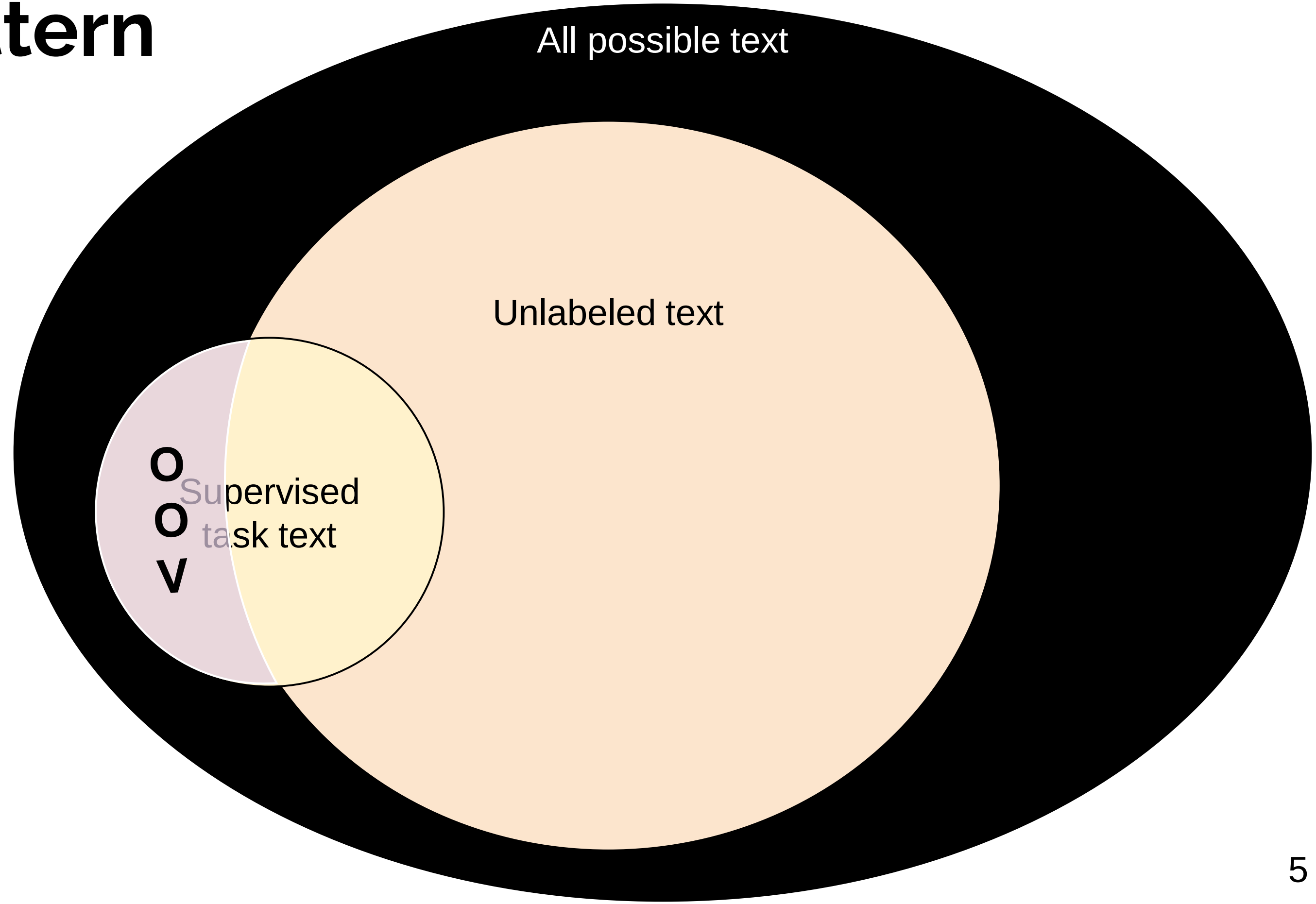
Assumed Pattern



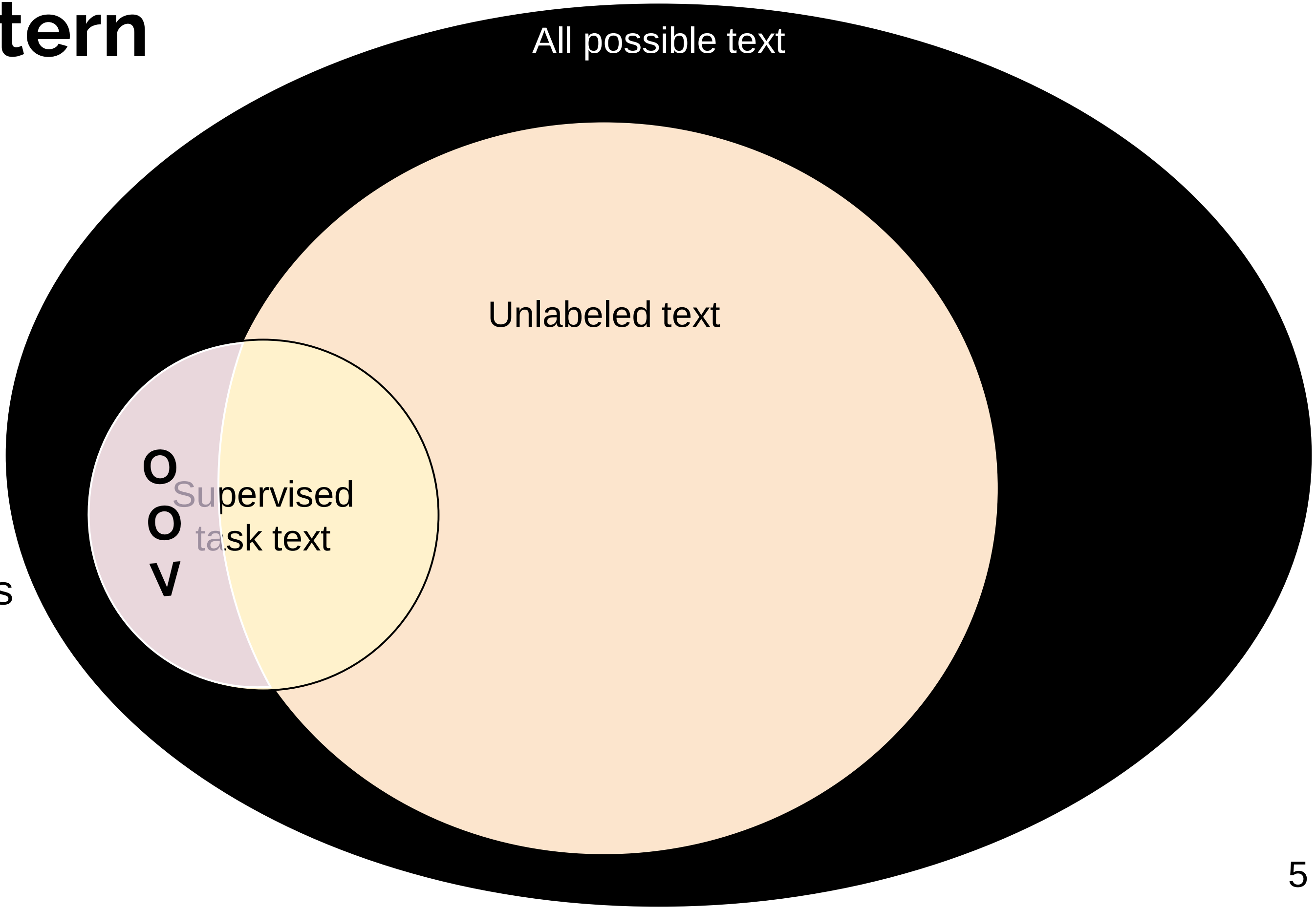
Actual Pattern



Actual Pattern

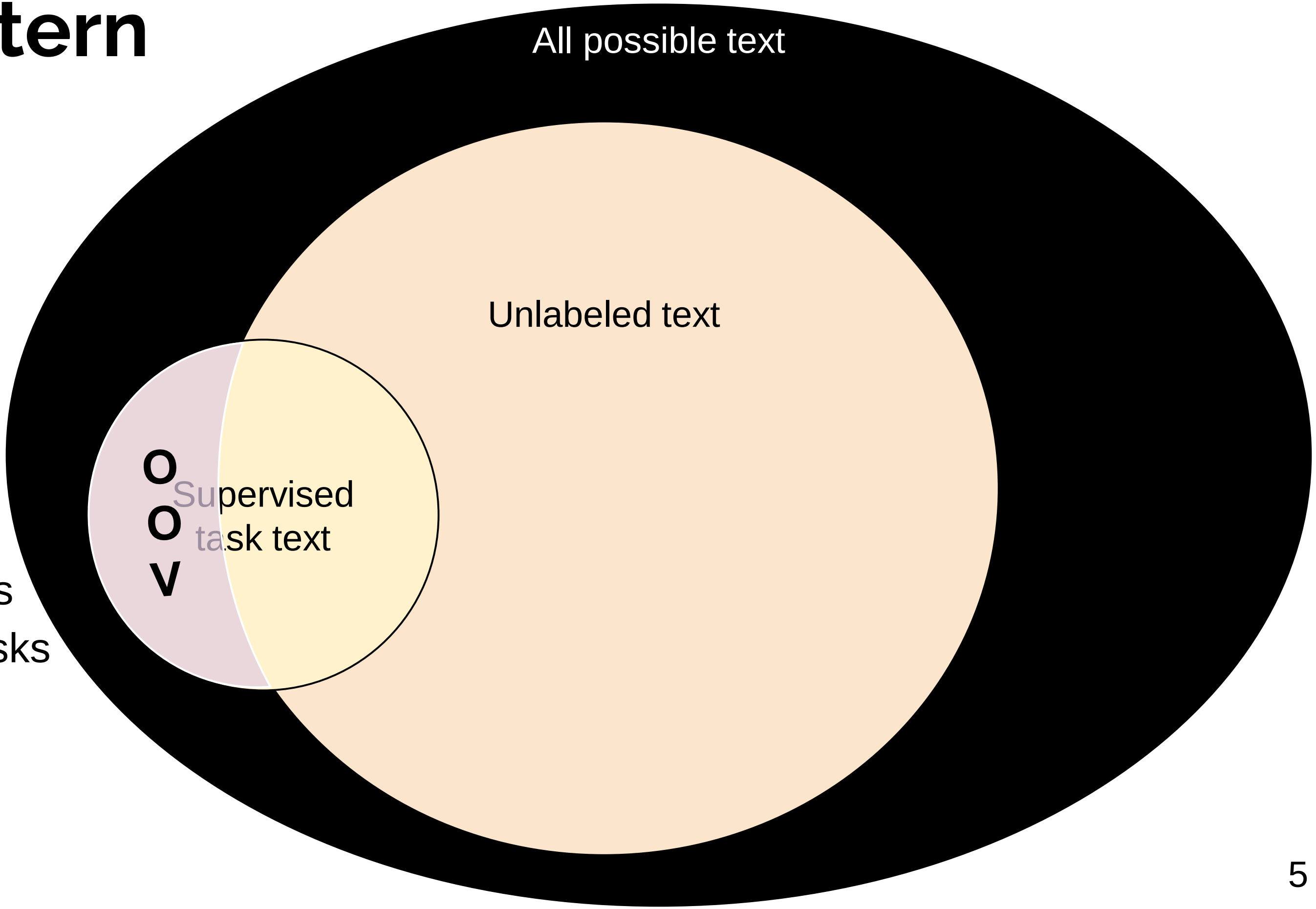


Actual Pattern



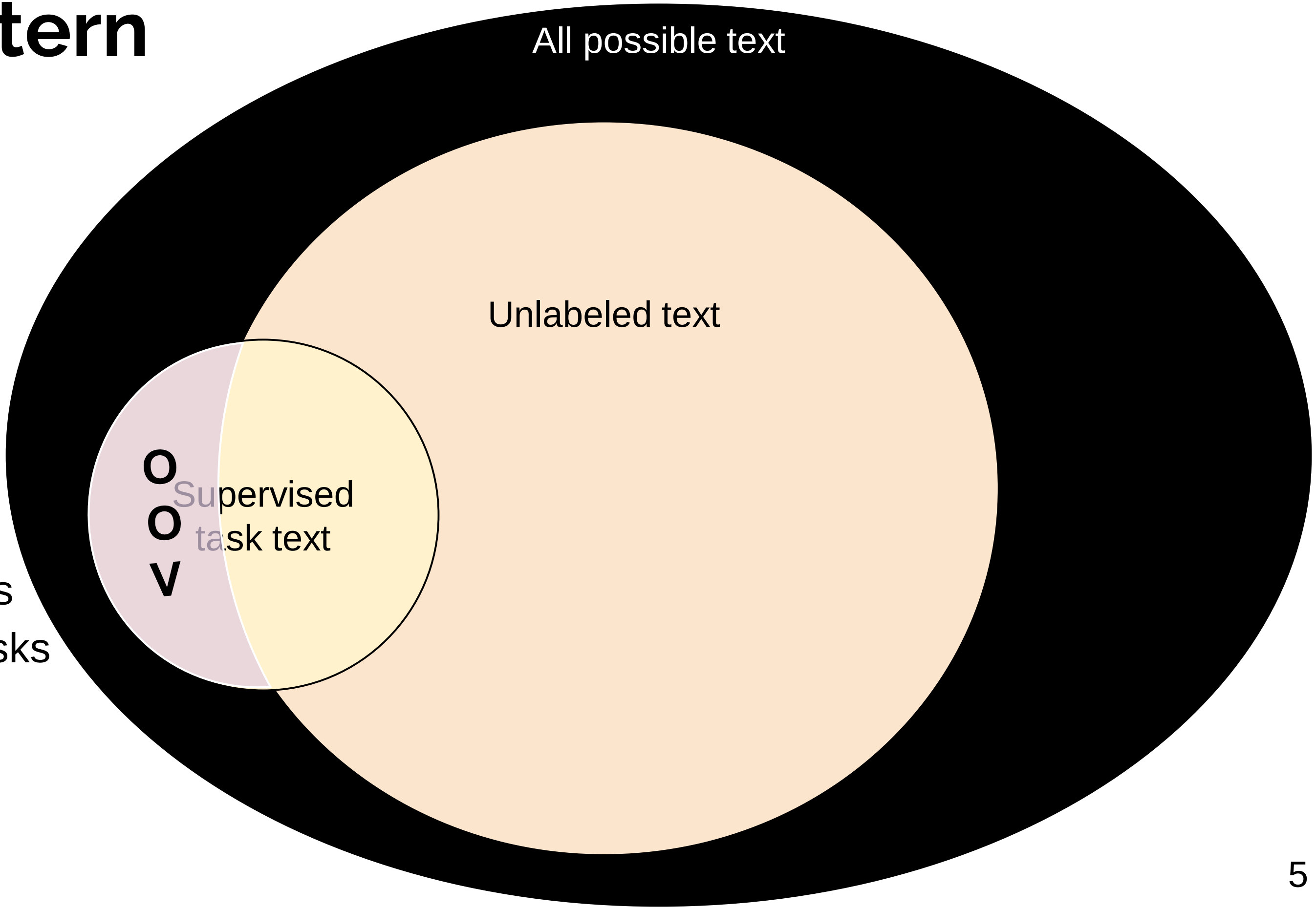
- No pre-trained vectors

Actual Pattern



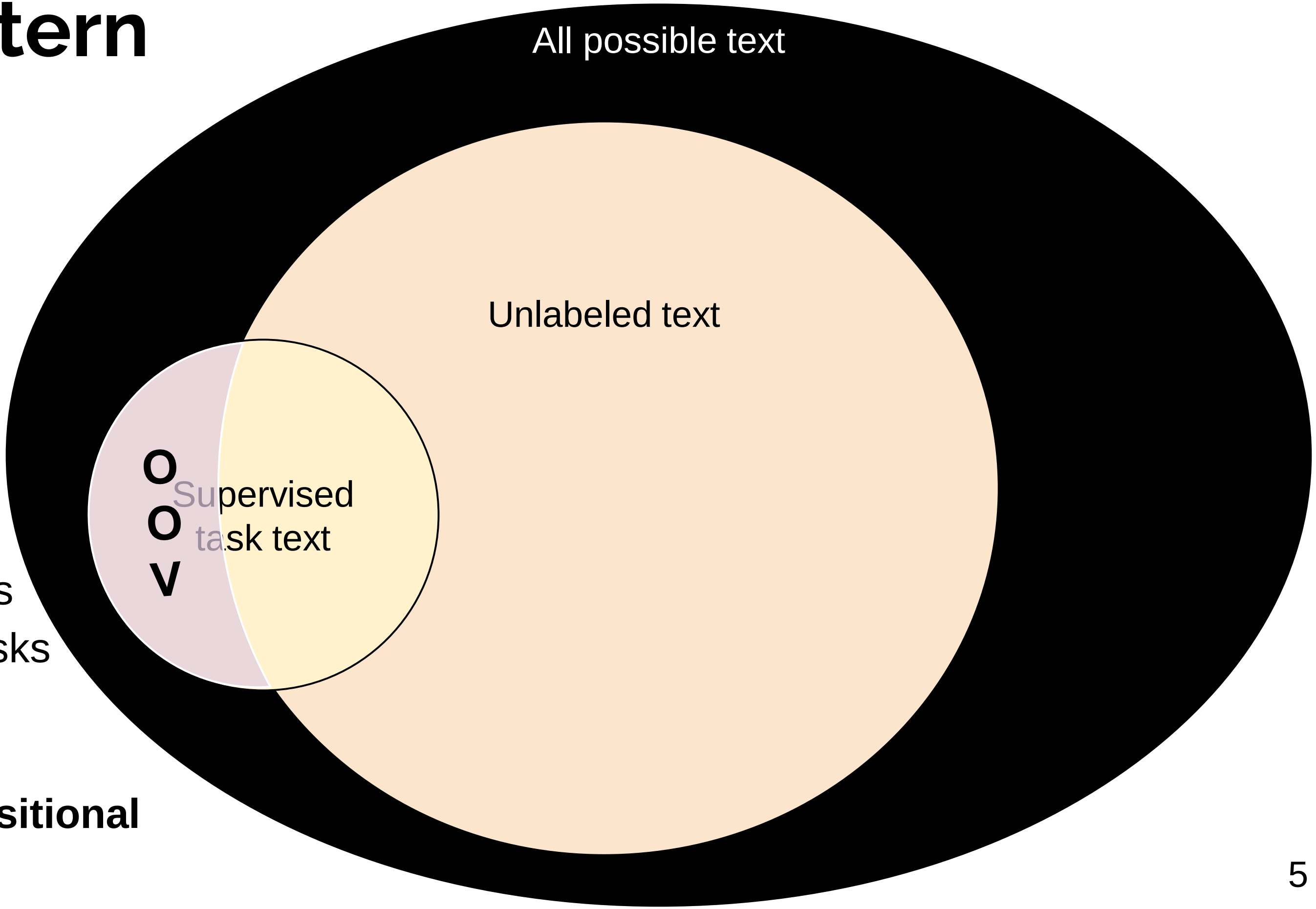
- No pre-trained vectors
- Affects supervised tasks

Actual Pattern



- No pre-trained vectors
- Affects supervised tasks
- Multiple treatments suggested

Actual Pattern



- No pre-trained vectors
- Affects supervised tasks
- Multiple treatments suggested
- Our method - **compositional** subword OOV model

Sources of OOVs

Sources of OOVs

- Names

Chalabi has increasingly marginalized within Iraq, ...

Sources of OOVs

- Names
- Domain-specific jargon

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words
- Rare morphological derivations

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Without George Martin the Beatles would have been just another **untalented** band as Oasis.

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words
- Rare morphological derivations
- Nonce words

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Without George Martin the Beatles would have been just another **untalented** band as Oasis.

What if Google morphed into **GoogleOS**?

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words
- Rare morphological derivations
- Nonce words
- Nonstandard orthography

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Without George Martin the Beatles would have been just another **untalented** band as Oasis.

What if Google morphed into **GoogleOS**?

We'll have four bands, and Big D is **cookin'**. lots of fun and great prizes.

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words
- Rare morphological derivations
- Nonce words
- Nonstandard orthography
- Typos and other errors

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Without George Martin the Beatles would have been just another **untalented** band as Oasis.

What if Google morphed into **GoogleOS**?

We'll have four bands, and Big D is **cookin'**. lots of fun and great prizes.

I dislike this urban society and I want to leave this whole **enviroment**.

Sources of OOVs

- Names
- Domain-specific jargon
- Foreign words
- Rare morphological derivations
- Nonce words
- Nonstandard orthography
- Typos and other errors
- …

Chalabi has increasingly marginalized within Iraq, ...

Important species (...) include shrimp, (...) and some varieties of **flatfish**.

This term was first used in German (**Hochrenaissance**), …

Without George Martin the Beatles would have been just another **untalented** band as Oasis.

What if Google morphed into **GoogleOS**?

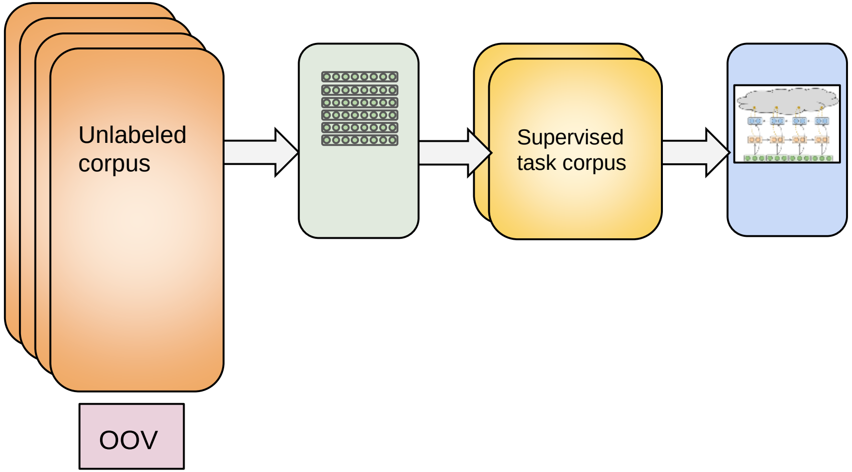
We'll have four bands, and Big D is **cookin'**. lots of fun and great prizes.

I dislike this urban society and I want to leave this whole **enviroment**.

???

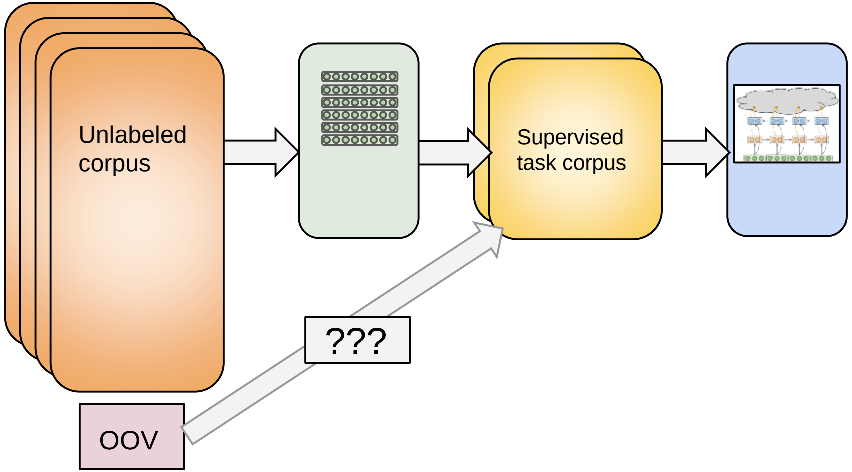
Common OOV handling techniques

- None (random init)



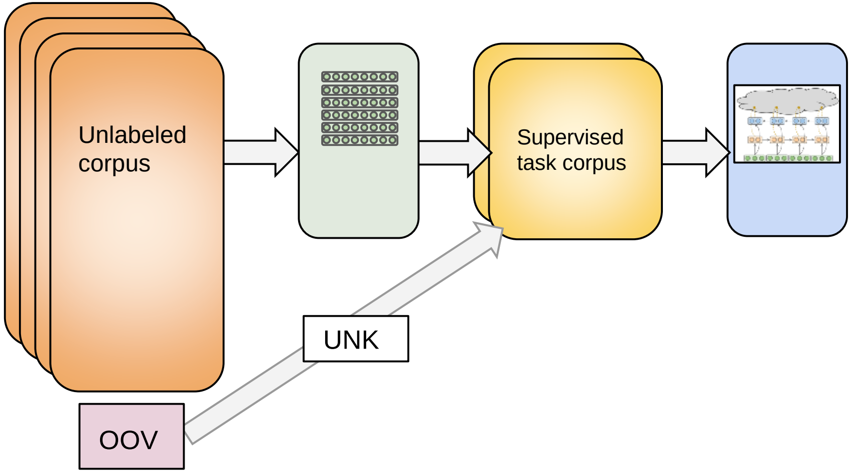
Common OOV handling techniques

- None (random init)



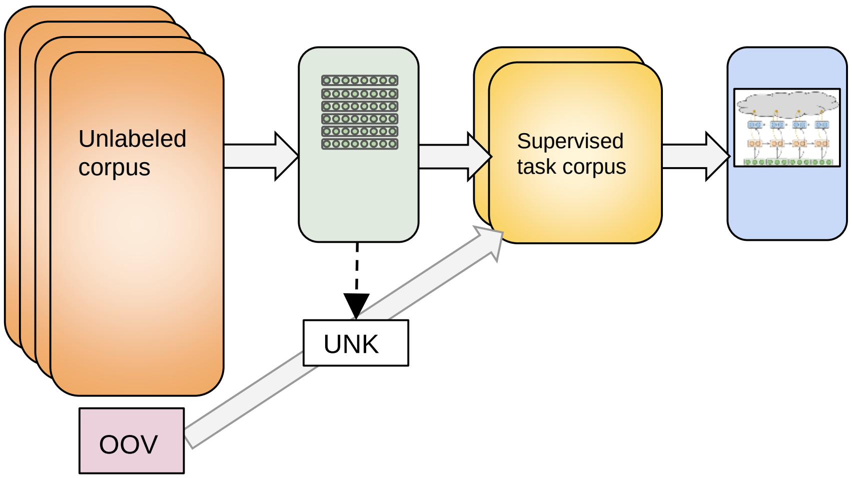
Common OOV handling techniques

- None (random init)
- One UNK to rule them all
  - Average existing embeddings
  - Trained with embeddings (stochastic unking)



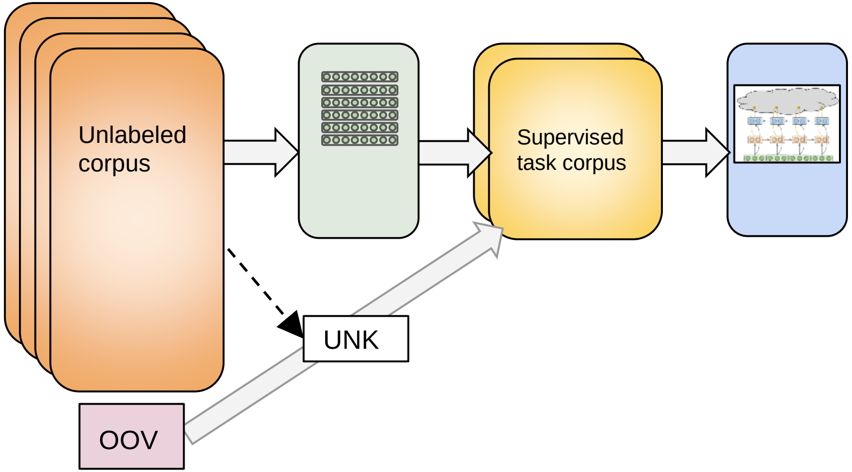
Common OOV handling techniques

- None (random init)
- One UNK to rule them all
    - Average existing embeddings
    - Trained with embeddings (stochastic unking)



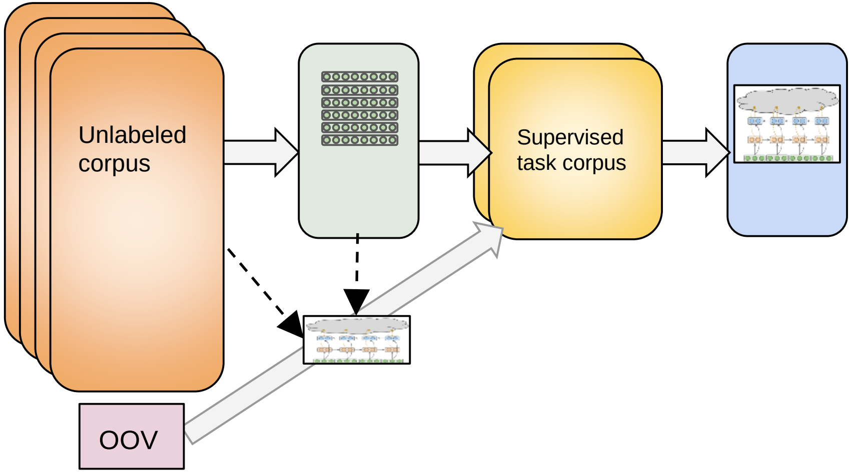
Common OOV handling techniques

- None (random init)
- One UNK to rule them all
    - Average existing embeddings
    - Trained with embeddings (stochastic unking)



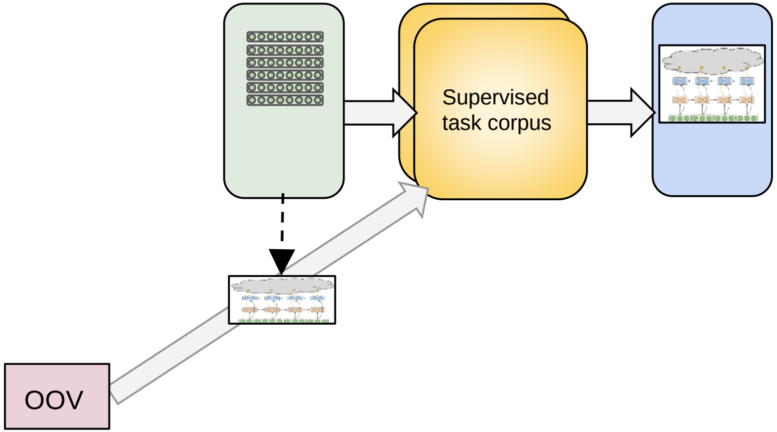
Common OOV handling techniques

- None (random init)
- One UNK to rule them all
  - Average existing embeddings
  - Trained with embeddings (stochastic unking)
- Add subword model during WE training
  - Bhatia et al. (2016), Wieting et al. (2016)

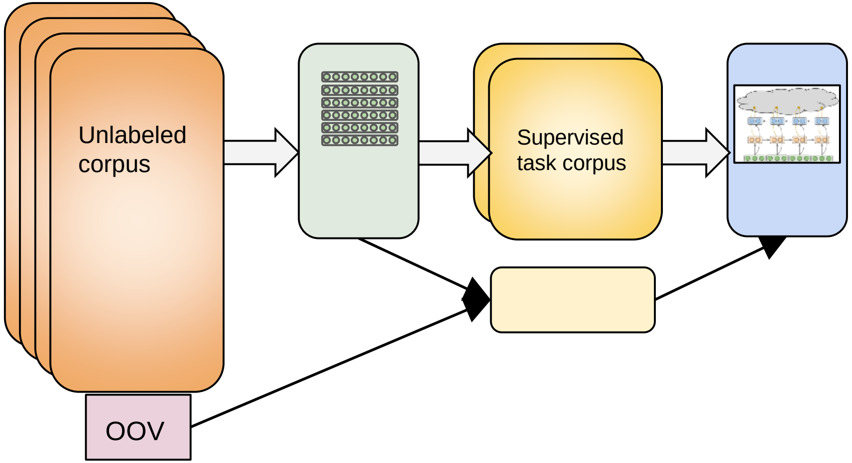


Common OOV handling techniques

- None (random init)
- One UNK to rule them all
    - Average existing embeddings
    - Trained with embeddings (stochastic unking)
- Add subword model during WE training
    - Bhatia et al. (2016), Wieting et al. (2016)
    - What if we don't have access to the original corpus? (e.g. FastText)

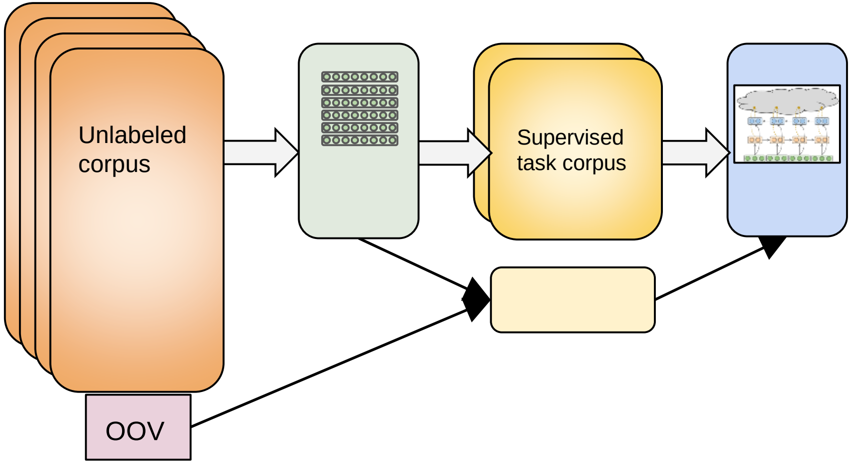


Char2Tag



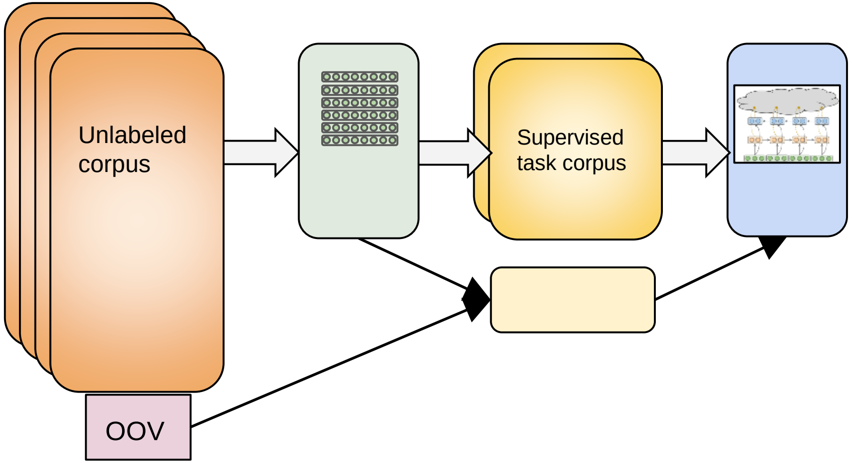
Char2Tag

- Add subword layer to supervised task
  - Ling et al. (2015), Plank et al. (2016)



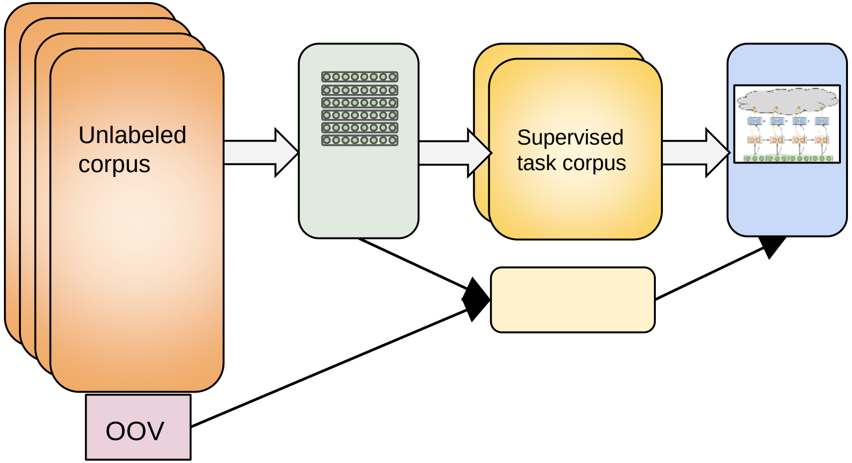
Char2Tag

- Add subword layer to supervised task
  - Ling et al. (2015), Plank et al. (2016)
- OOVs benefit from co-trained character model

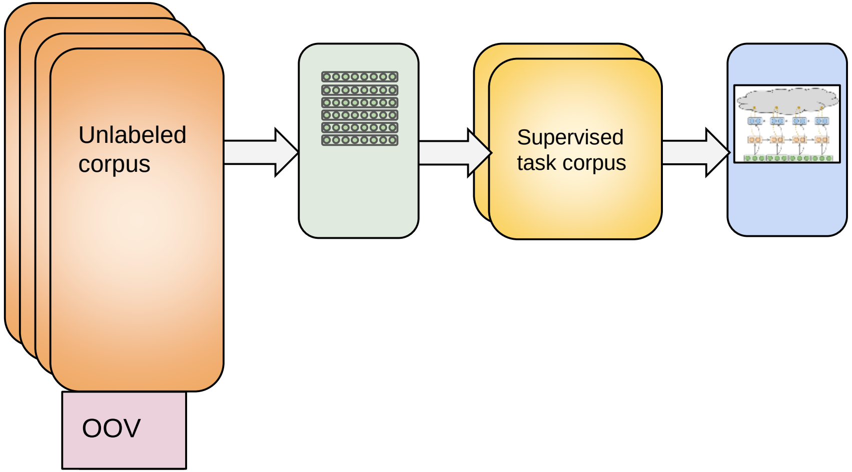


Char2Tag

- Add subword layer to supervised task
  - Ling et al. (2015), Plank et al. (2016)
- OOVs benefit from co-trained character model
- Requires large supervised training set for efficient transfer to test set OOVs

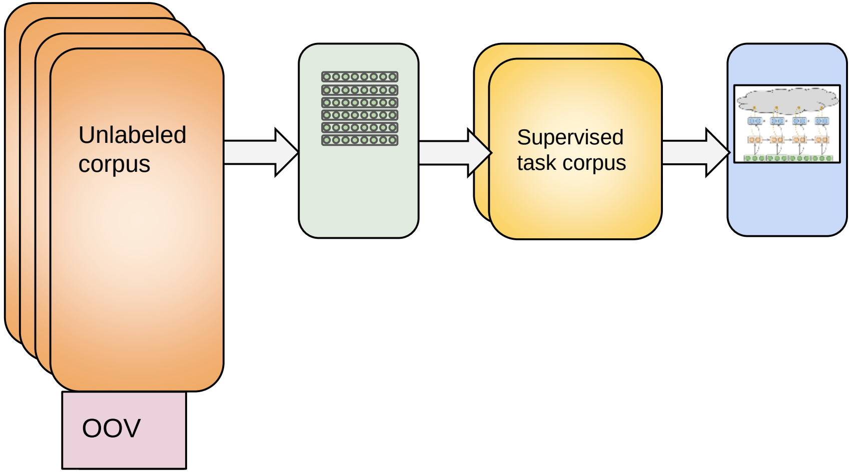


Enter MIMICK



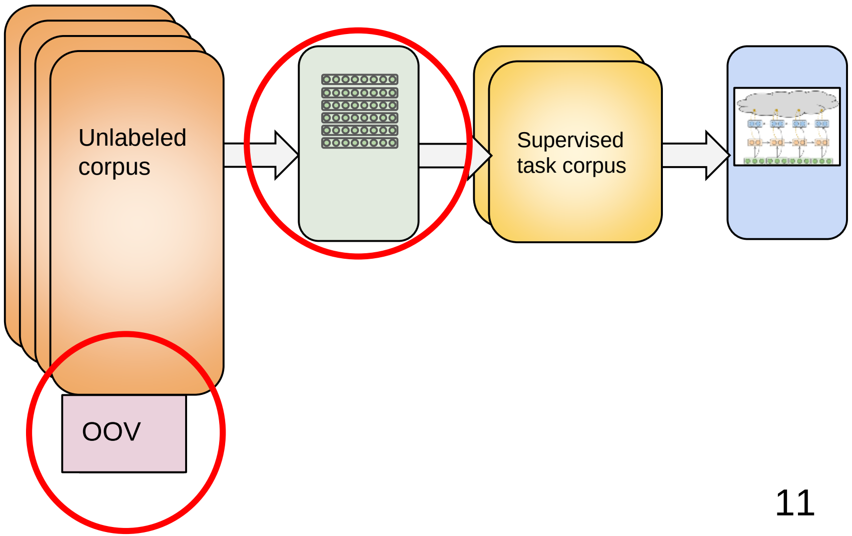
Enter MIMICK

- What data do we have, post-unlabeled corpus?
  - Vector dictionary
  - Orthography (the way words are spelled)



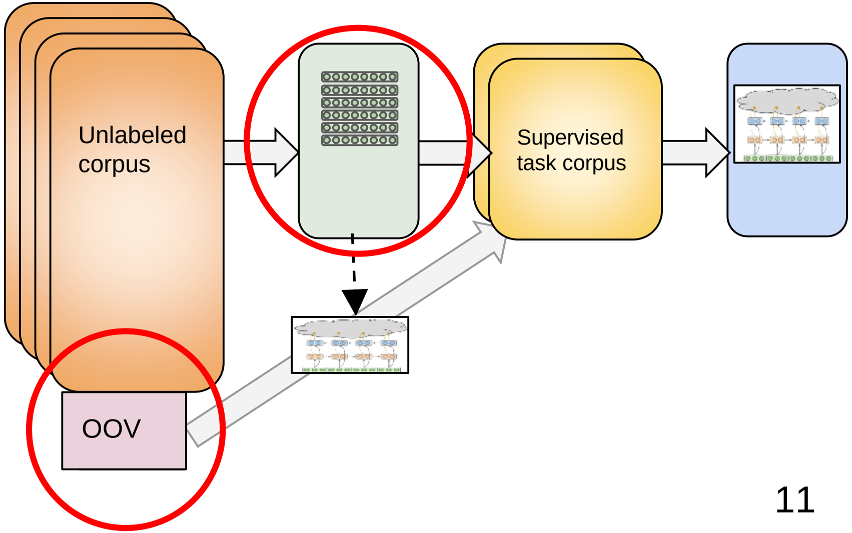
Enter MIMICK

- What data do we have, post-unlabeled corpus?
  - Vector dictionary
  - Orthography (the way words are spelled)



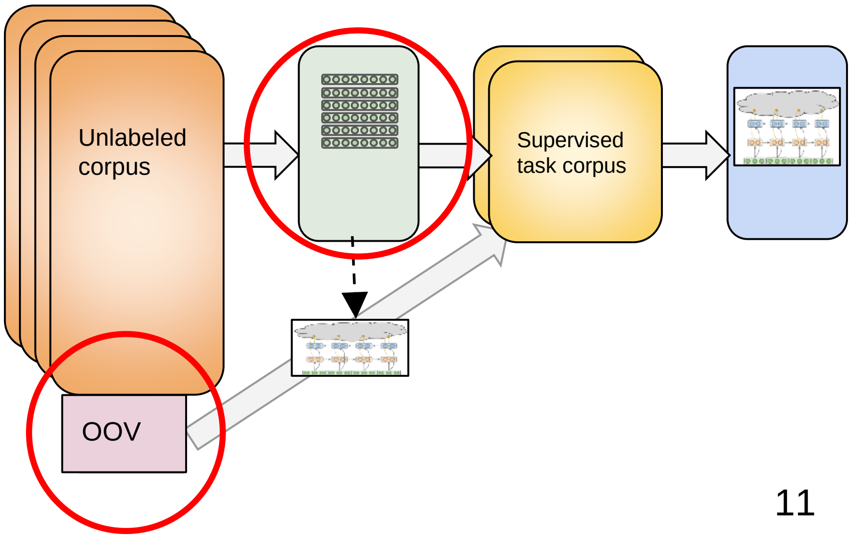
Enter MIMICK

- What data do we have, post-unlabeled corpus?
  - Vector dictionary
  - Orthography (the way words are spelled)
- **Use the former as training objective, latter as input**



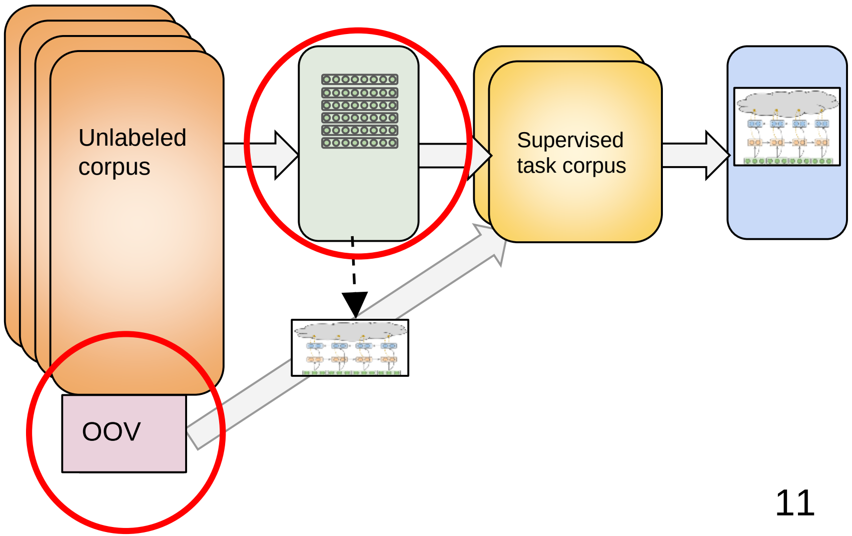
Enter MIMICK

- What data do we have, post-unlabeled corpus?
  - Vector dictionary
  - Orthography (the way words are spelled)
- **Use the former as training objective, latter as input**
- Pre-trained vectors as target
  - No need to access original unlabeled corpus
  - Many training examples
  - (No context)



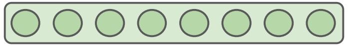
Enter MIMICK

- What data do we have, post-unlabeled corpus?
  - Vector dictionary
  - Orthography (the way words are spelled)
- **Use the former as training objective, latter as input**
- Pre-trained vectors as target
  - No need to access original unlabeled corpus
  - Many training examples
  - (No context)
- Subword units as inputs
  - Very extensible
  - (Character inventory changes?)

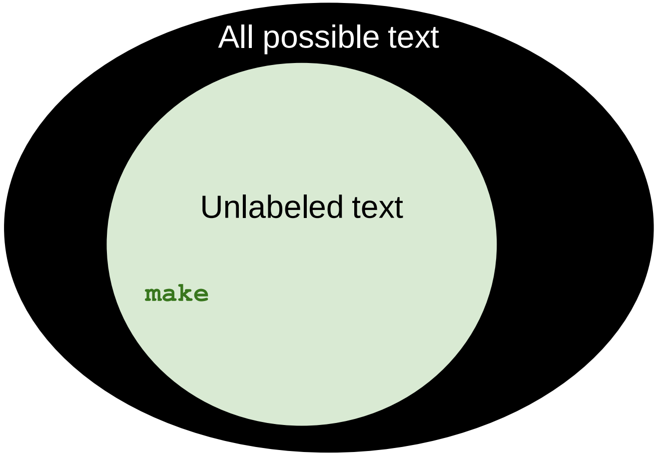


MIMICK Training

Pre-trained Embedding
(Polyglot/FastText/etc.)

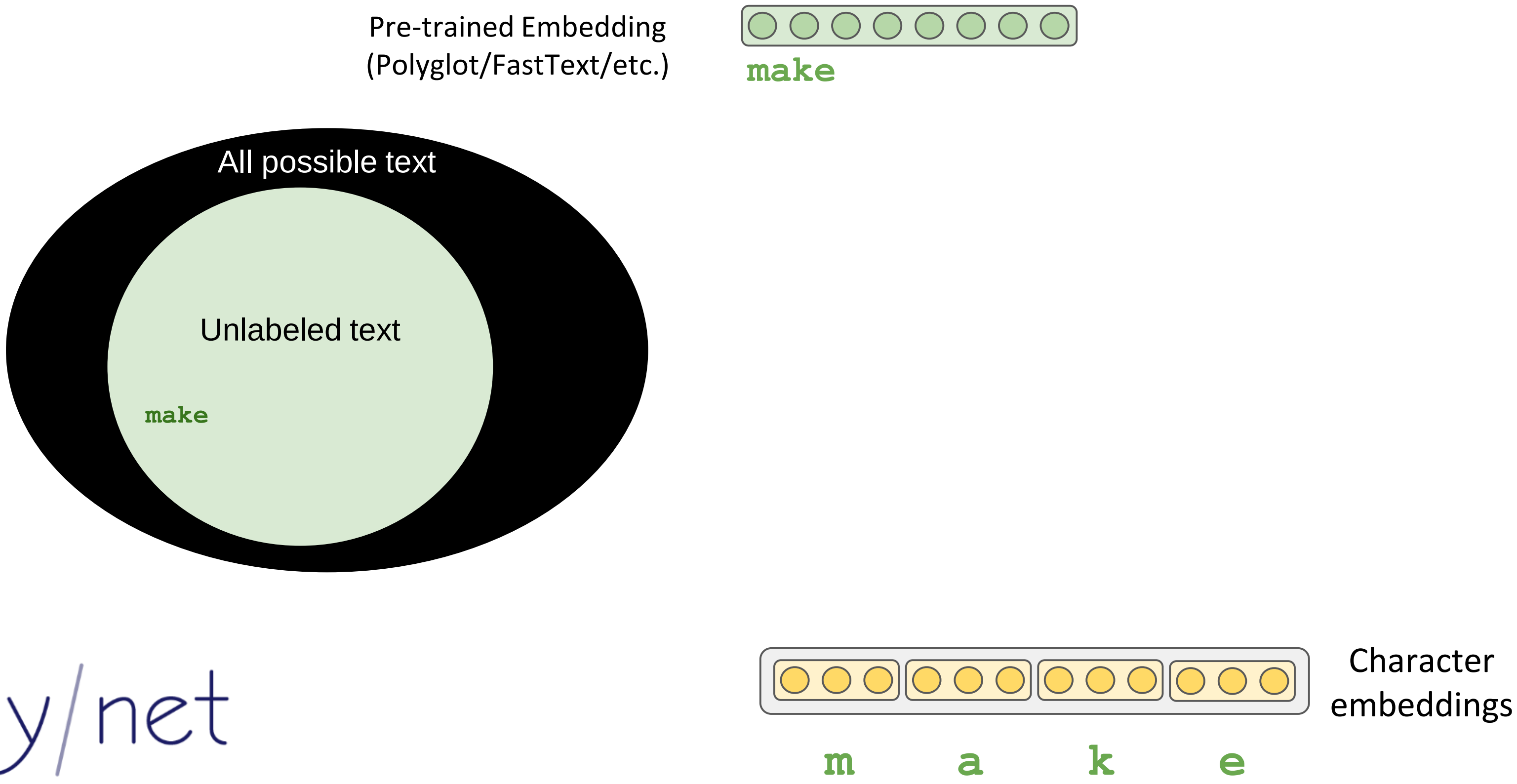


make

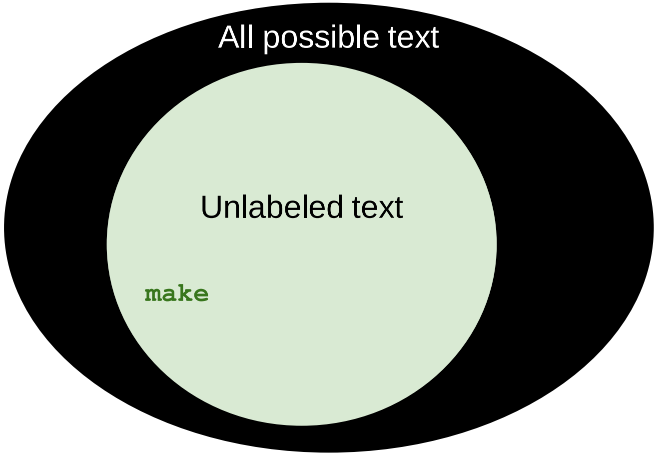


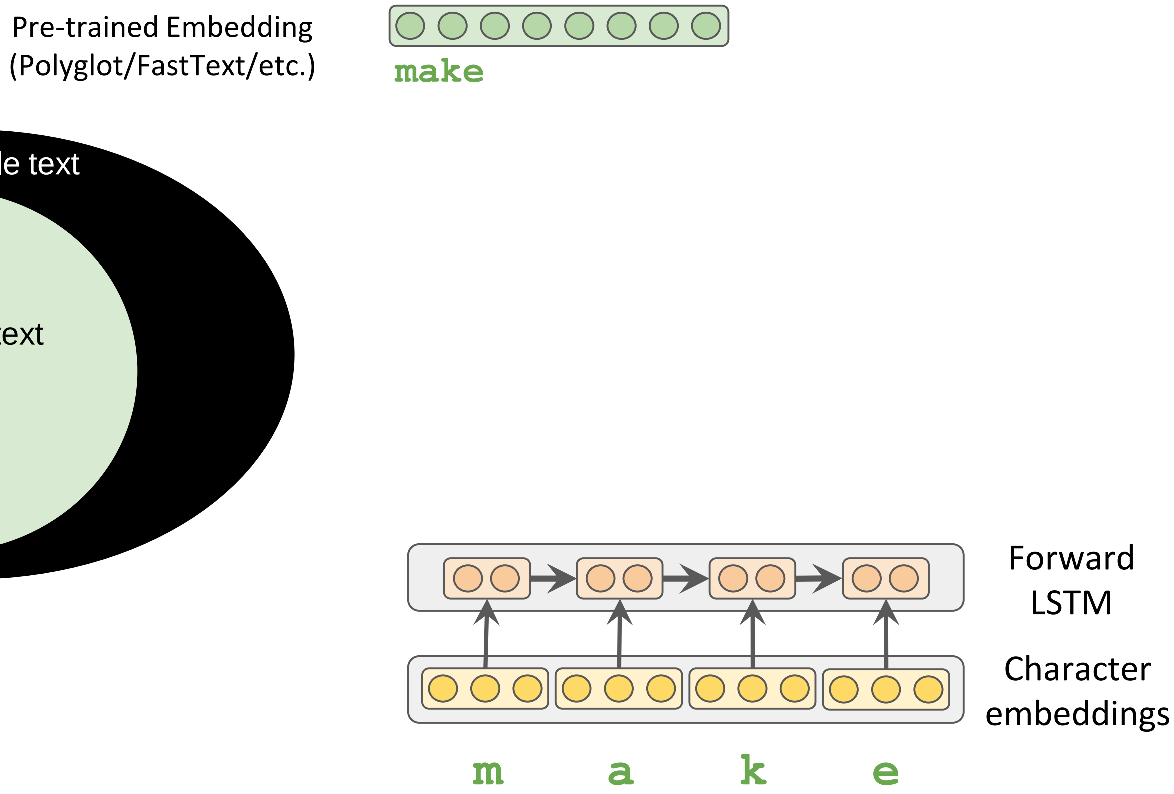
m a k e

MIMICK Training

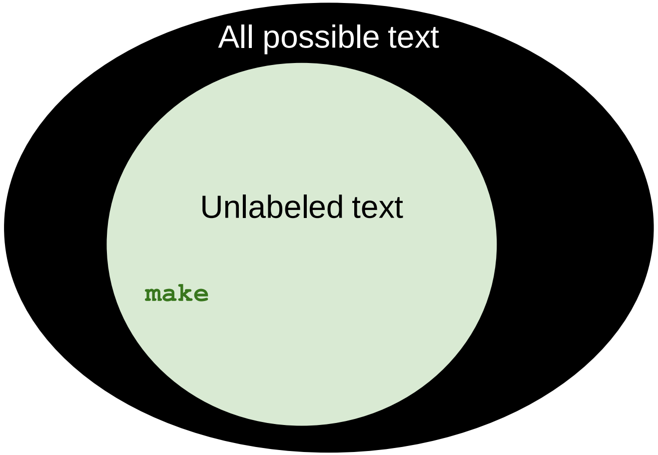


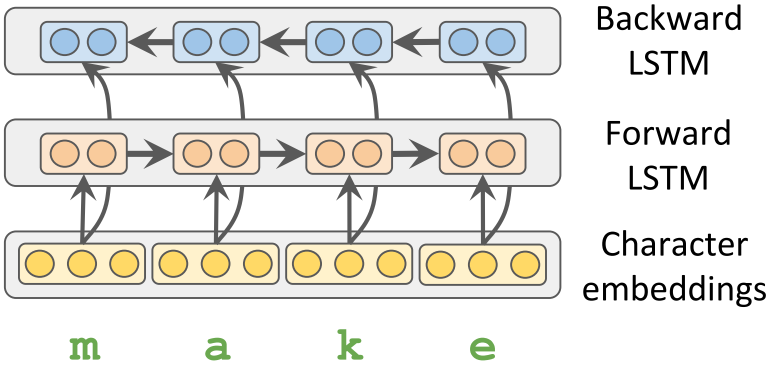
MIMICK Training





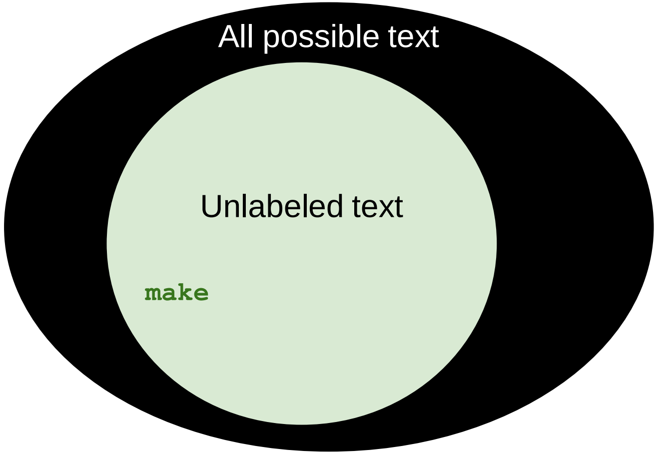
MIMICK Training

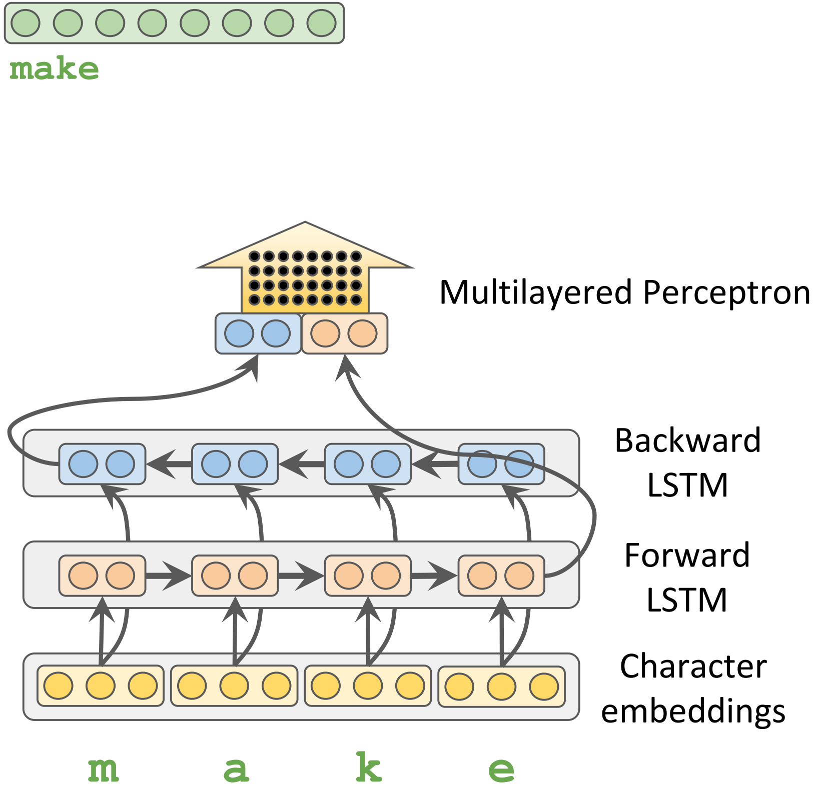




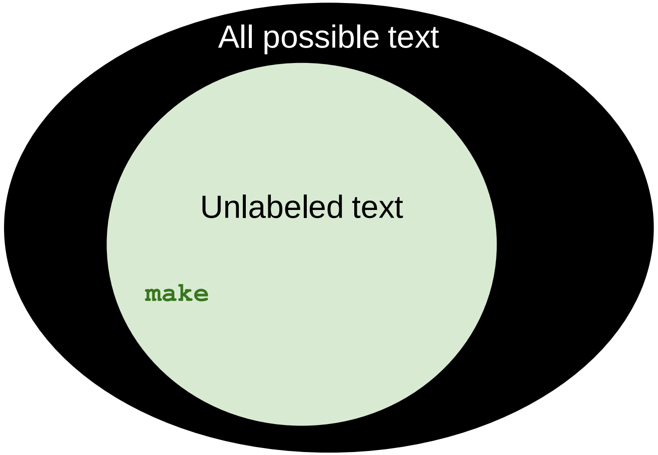
MIMICK Training

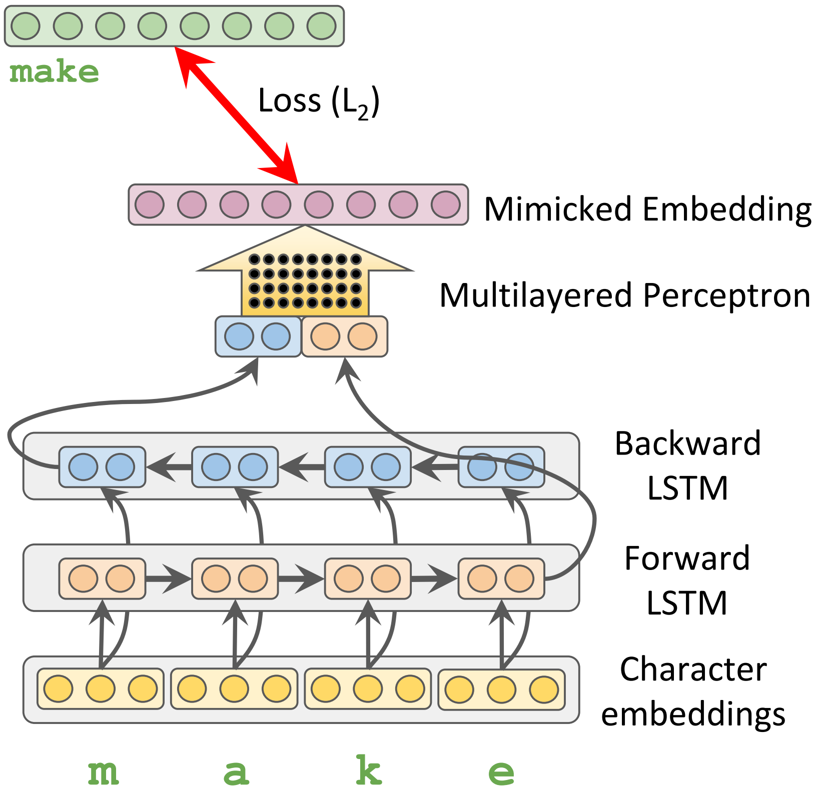
Pre-trained Embedding
(Polyglot/FastText/etc.)



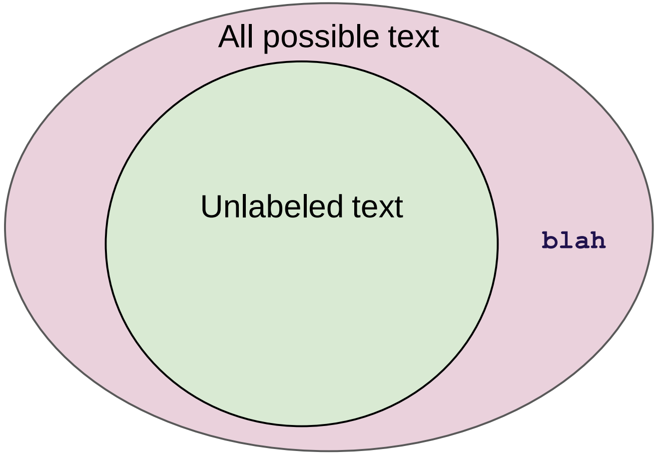


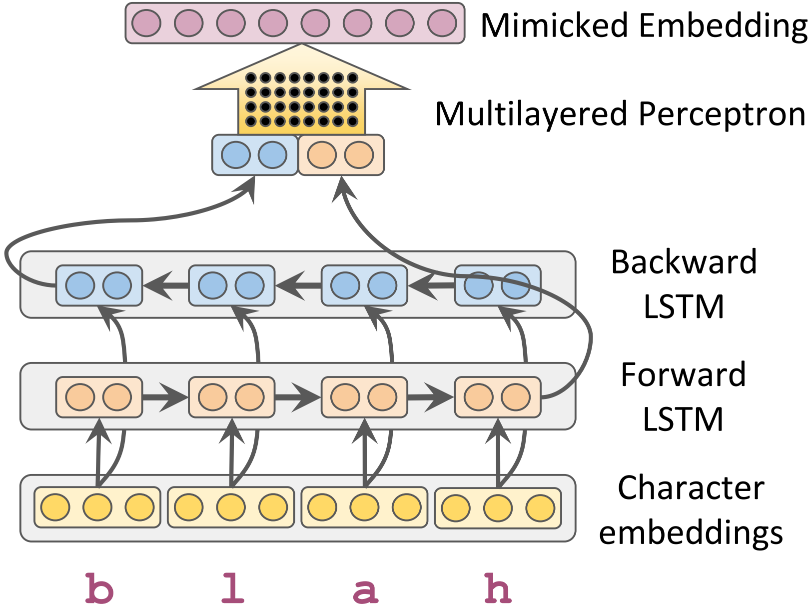
MIMICK Training





MIMICK Inference





Observation – Nearest Neighbors

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
    - MCT → AWS, OTA, APT, PDM

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
    - MCT → AWS, OTA, APT, PDM
    - pesky → euphoric, disagreeable, horrid, ghastly

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
  - MCT → AWS, OTA, APT, PDM
  - pesky → euphoric, disagreeable, horrid, ghastly
  - lawnmower → tradesman, bookmaker, postman, hairdresser

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
  - MCT → AWS, OTA, APT, PDM
  - pesky → euphoric, disagreeable, horrid, ghastly
  - lawnmower → tradesman, bookmaker, postman, hairdresser

- Hebrew

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
  - MCT → AWS, OTA, APT, PDM
  - pesky → euphoric, disagreeable, horrid, ghastly
  - lawnmower → tradesman, bookmaker, postman, hairdresser

- Hebrew
  - תתגשם → תפתור (she/you-3p.sg.) will come true (she/you-3p.sg.) will solve

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
    - MCT → AWS, OTA, APT, PDM
    - pesky → euphoric, disagreeable, horrid, ghastly
    - lawnmower → tradesman, bookmaker, postman, hairdresser

- Hebrew

○ תתגשם ⟶ תפתור	(she/you-3p.sg.) will come true	(she/you-3p.sg.) will solve
○ גיאומטריים ⟶ גאומטריים	geometric (m.pl., nontrad. spelling)	geometric (m.pl.)

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
  - MCT → AWS, OTA, APT, PDM
  - pesky → euphoric, disagreeable, horrid, ghastly
  - lawnmower → tradesman, bookmaker, postman, hairdresser

- Hebrew

○ תתגשם → תפתור	(she/you-3p.sg.) will come true	(she/you-3p.sg.) will solve
○ גיאומטריים → גאומטריים	geometric (m.pl., nontrad. spelling)	geometric (m.pl.)
○ ריצ’רדסון → אויסטרך	Richardson	Eustrach

Observation – Nearest Neighbors

- English (OOV → Nearest in-vocab words)
  - MCT → AWS, OTA, APT, PDM
  - pesky → euphoric, disagreeable, horrid, ghastly
  - lawnmower → tradesman, bookmaker, postman, hairdresser

- Hebrew

תתגשם ⟶ תפתור	(she/you-3p.sg.) will come true	(she/you-3p.sg.) will solve
גיאומטריים ⟶ גאומטריים	geometric (m.pl., nontrad. spelling)	geometric (m.pl.)
ריצ’רדסון ⟶ אויסטרך	Richardson	Eustrach

- ✓ Surface form ✓ Syntactic properties ✗ Semantics

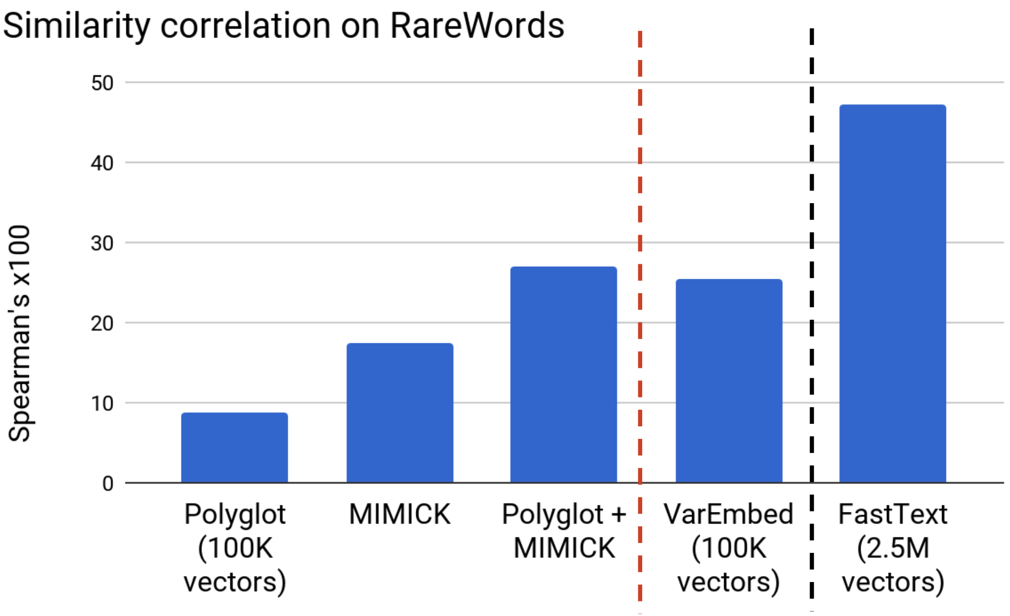
Intrinsic Evaluation – RareWords

Intrinsic Evaluation – RareWords

- RareWords similarity task: morphologically-complex, mostly unseen words

Intrinsic Evaluation – RareWords

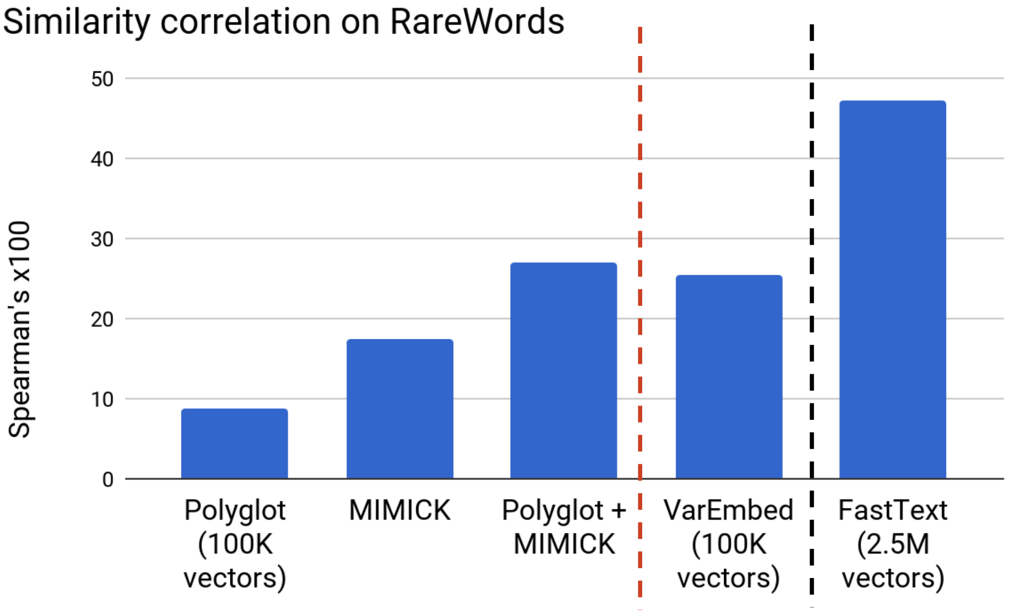
- RareWords similarity task: morphologically-complex, mostly unseen words



Intrinsic Evaluation – RareWords

- RareWords similarity task: morphologically-complex, mostly unseen words

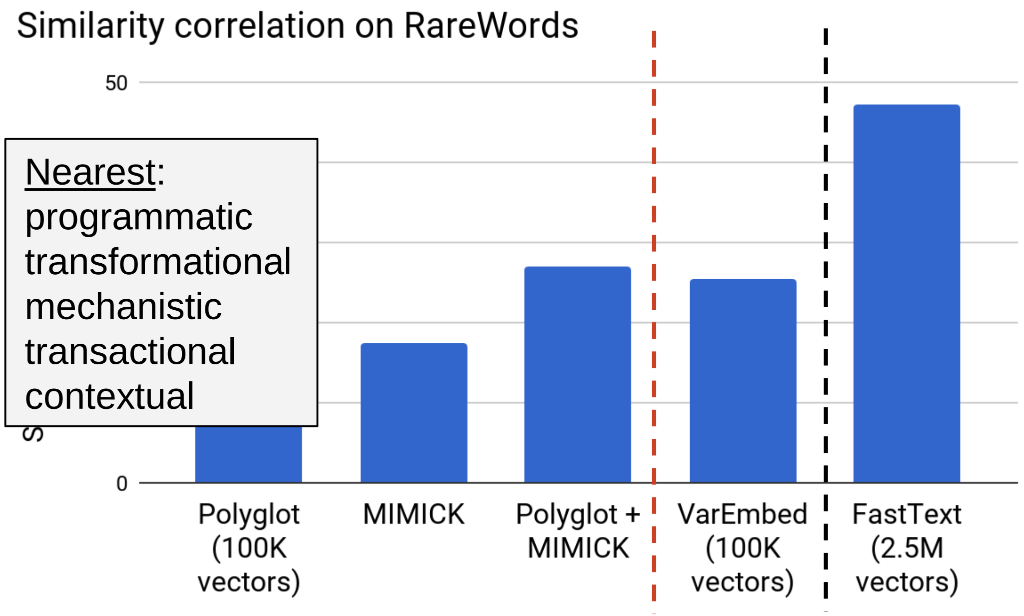
- Names
- Domain-specific jargon
- Foreign words
- **Rare(-ish) morphological derivations**
- Nonstandard orthography
- Typos and other errors
- ...



Intrinsic Evaluation – RareWords

- RareWords similarity task: morphologically-complex, mostly unseen words

- Names
- Domain-specific jargon
- Foreign words
- **Rare(-ish) morphological derivations**
- Nonce words
- Nonstandard orthography
- Typos and other errors
- ...



Nearest:
programmatic
transformational
mechanistic
transactional
contextual

Extrinsic Evaluation – POS + Attribute Tagging

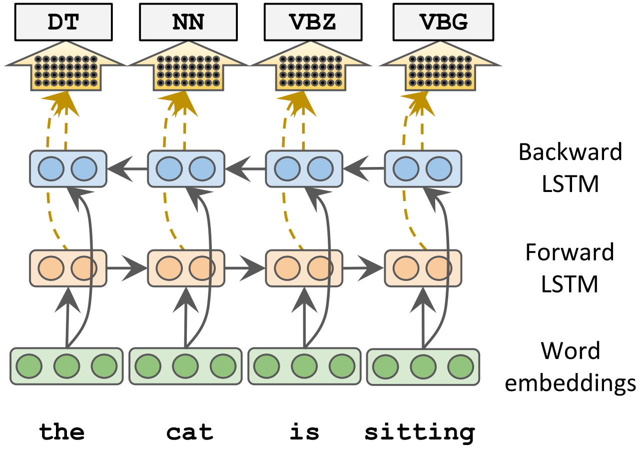
- UD is annotated for POS and morphosyntactic attributes
  - Eng: his **stated** goals — Tense=Past|VerbForm=Part
  - Cze: osoby v **pokročilém** věku — Animacy=Inan|Case=Loc|Degree=Pos|Gender=Masc|Negative=Pos|Number=Sing
    people of **advanced** age

- **Names**
- **Domain-specific jargon**
- **Foreign words**
- **Rare(-ish) morphological derivations**
- **Nonce words**
- **Nonstandard orthography**
- **Typos and other errors**
- …

Extrinsic Evaluation – POS + Attribute Tagging

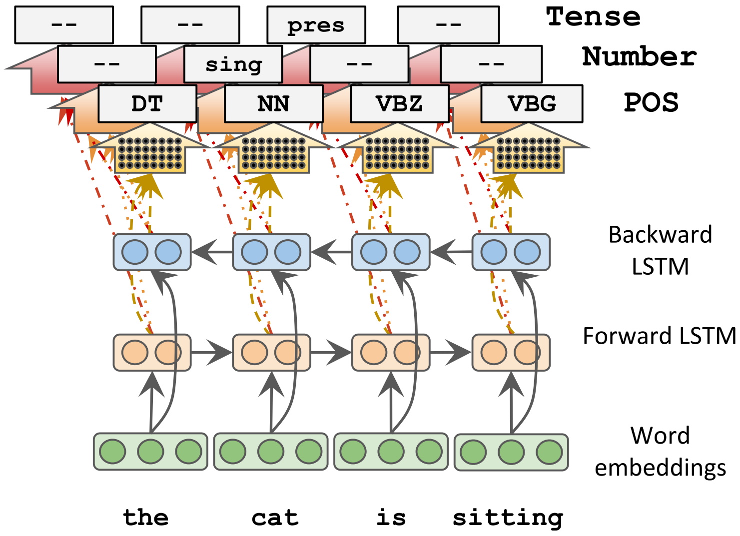
- UD is annotated for POS and morphosyntactic attributes
  - Eng: his **stated** goals — Tense=Past|VerbForm=Part
  - Cze: osoby v **pokročilém** věku (people of **advanced** age) — Animacy=Inan|Case=Loc|Degree=Pos|Gender=Masc|Negative=Pos|Number=Sing
- POS model from Ling et al. (2015)

- Names
- Domain-specific jargon
- Foreign words
- Rare(-ish) morphological derivations
- Nonce words
- Nonstandard orthography
- Typos and other errors
- …



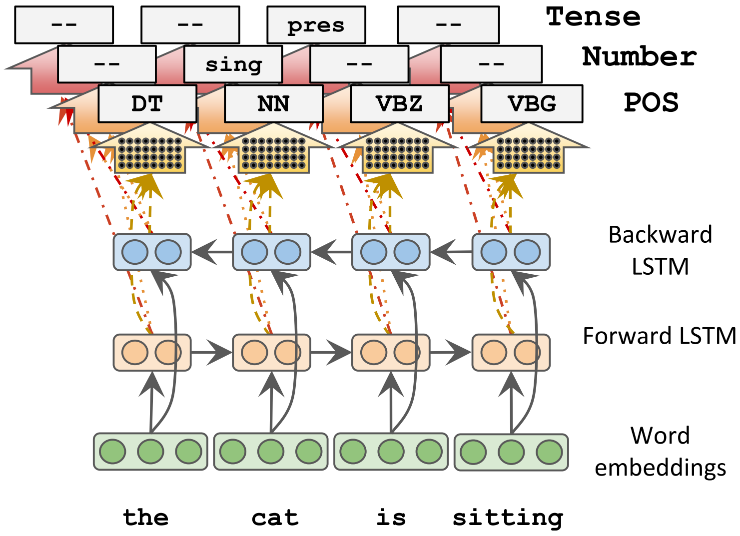
Extrinsic Evaluation – POS + Attribute Tagging

- UD is annotated for POS and morphosyntactic attributes
  - Eng: his **stated** goals — Tense=Past|VerbForm=Part
  - Cze: osoby v **pokročilém** věku — Animacy=Inan|Case=Loc|Degree=Pos|Gender=Masc|Negative=Pos|Number=Sing
    people of **advanced** age
- POS model from Ling et al. (2015)
- Attributes - same as POS layer



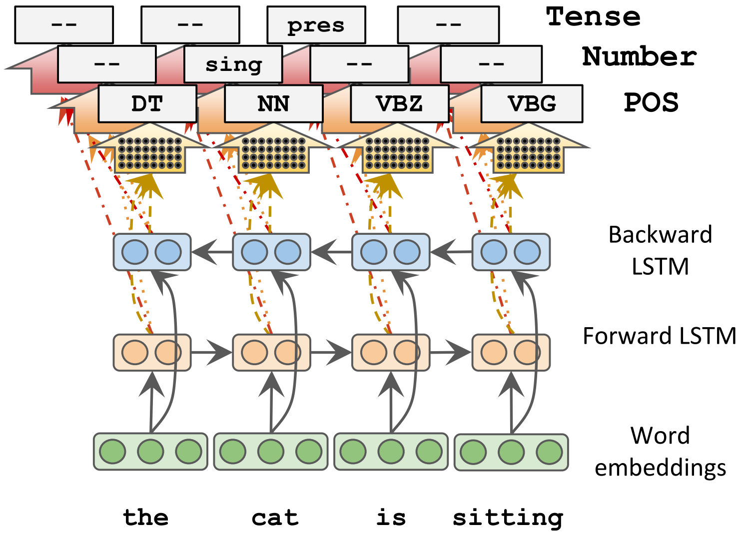
Extrinsic Evaluation – POS + Attribute Tagging

- UD is annotated for POS and morphosyntactic attributes
  - Eng: his **stated** goals — Tense=Past|VerbForm=Part
  - Cze: osoby v **pokročilém** věku — Animacy=Inan|Case=Loc|Degree=Pos|Gender=Masc|Negative=Pos|Number=Sing
    people of **advanced** age
- POS model from Ling et al. (2015)
- Attributes - same as POS layer
- Negative effect on POS

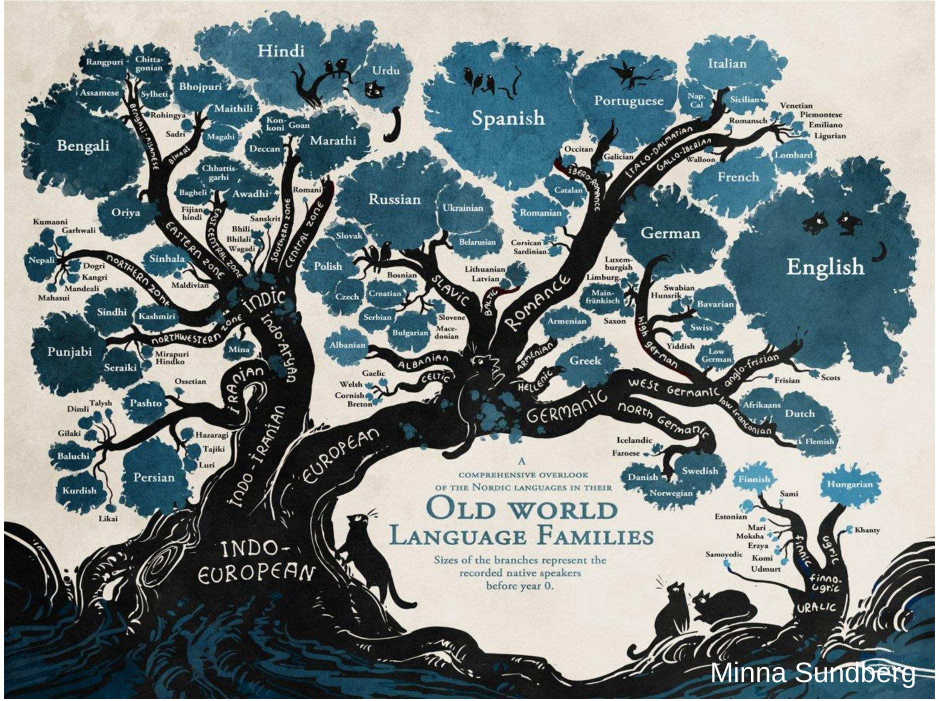


Extrinsic Evaluation – POS + Attribute Tagging

- UD is annotated for POS and morphosyntactic attributes
  - Eng: his **stated** goals — Tense=Past|VerbForm=Part
  - Cze: osoby v **pokročilém** věku — Animacy=Inan|Case=Loc|Degree=Pos|Gender=Masc|Negative=Pos|Number=Sing
    people of **advanced** age
- POS model from Ling et al. (2015)
- Attributes - same as POS layer
- Negative effect on POS
- Attribute evaluation metric
  - Micro F1



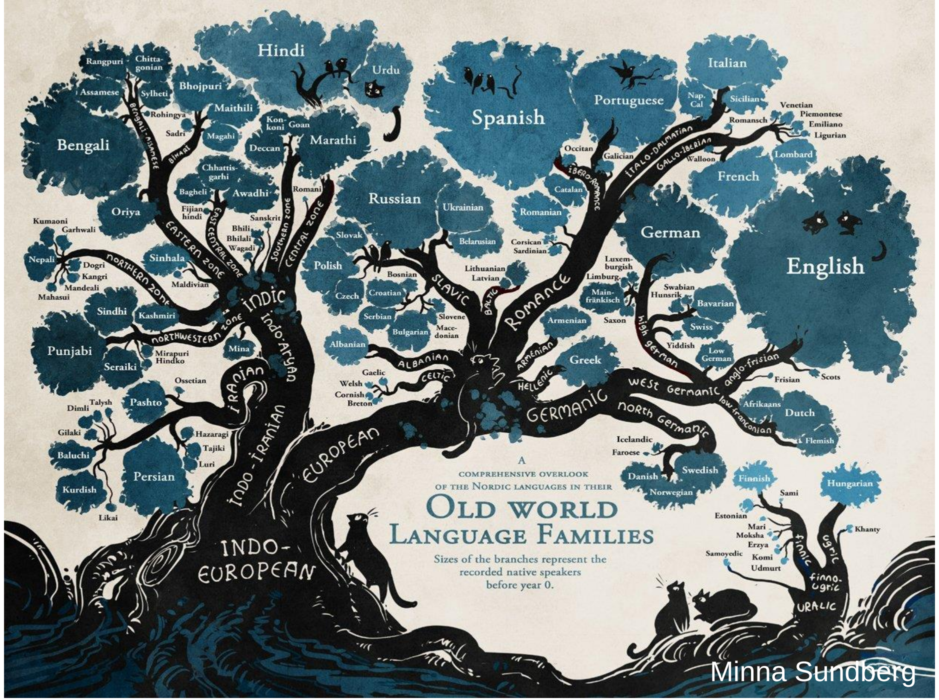
Language Selection



Minna Sundberg

Language Selection

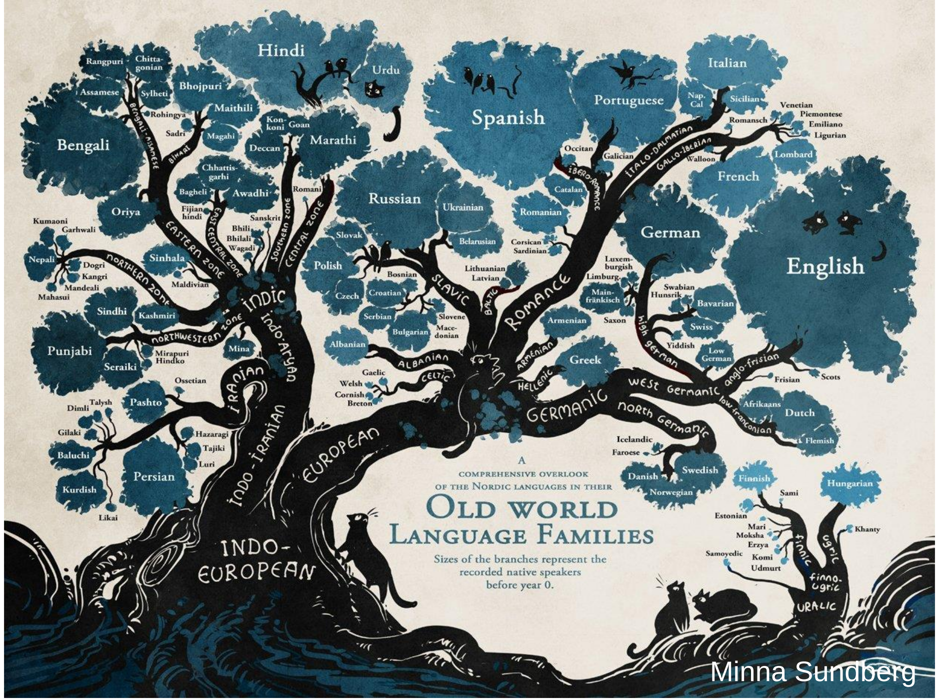
- |UD ∩ Polyglot| = 44, we took 23



Minna Sundberg

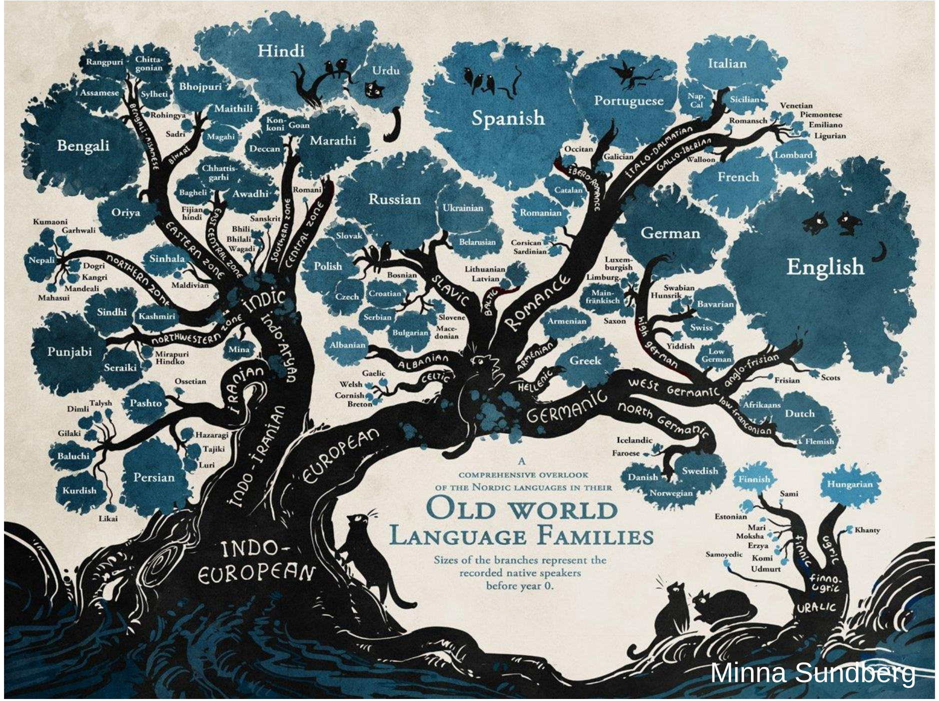
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure



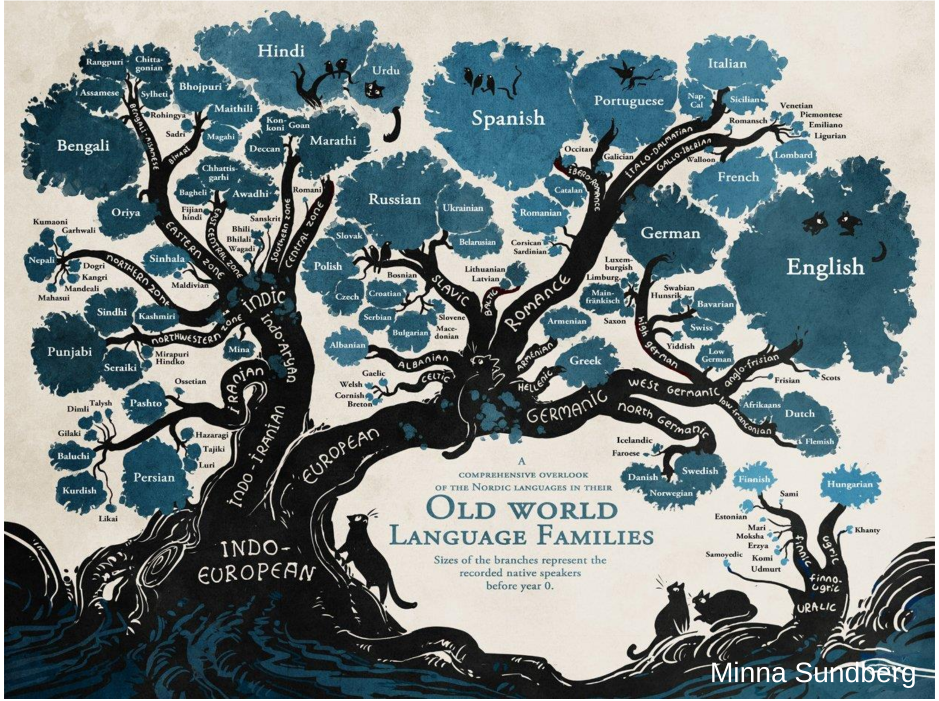
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional



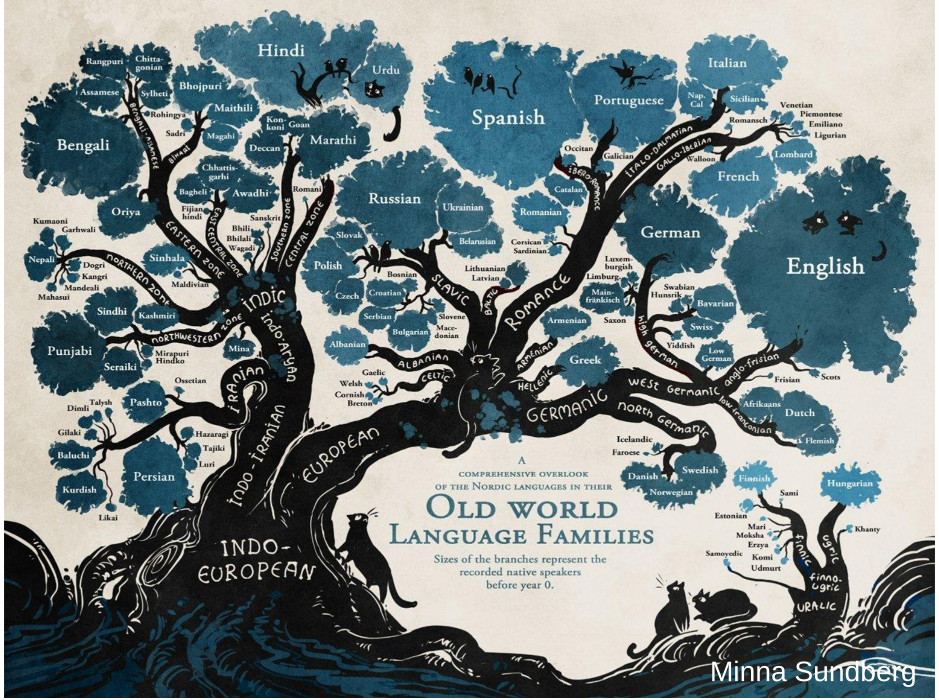
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic



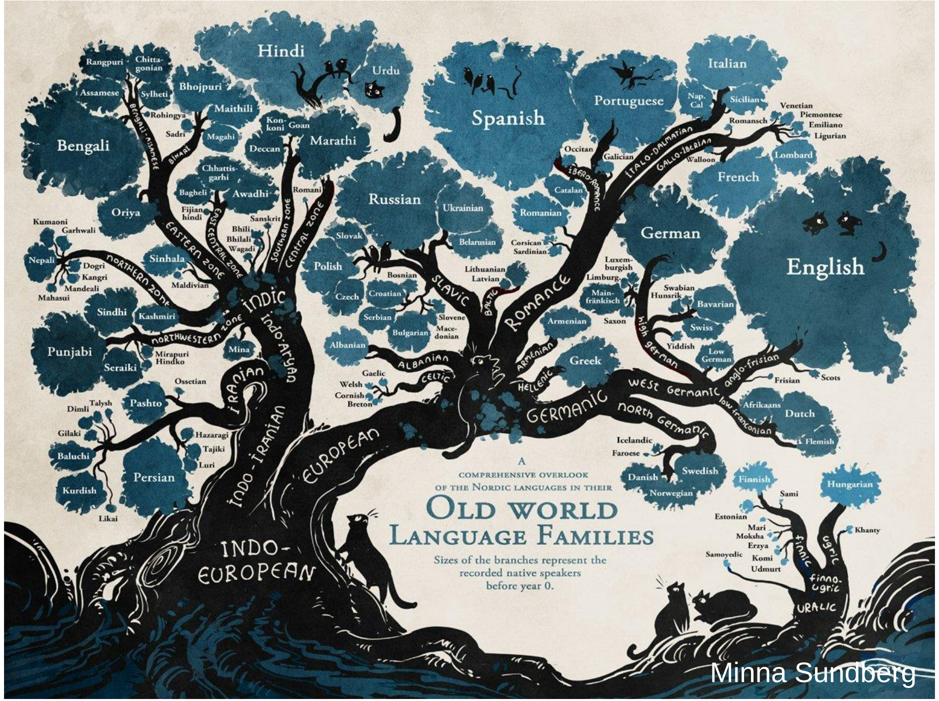
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating



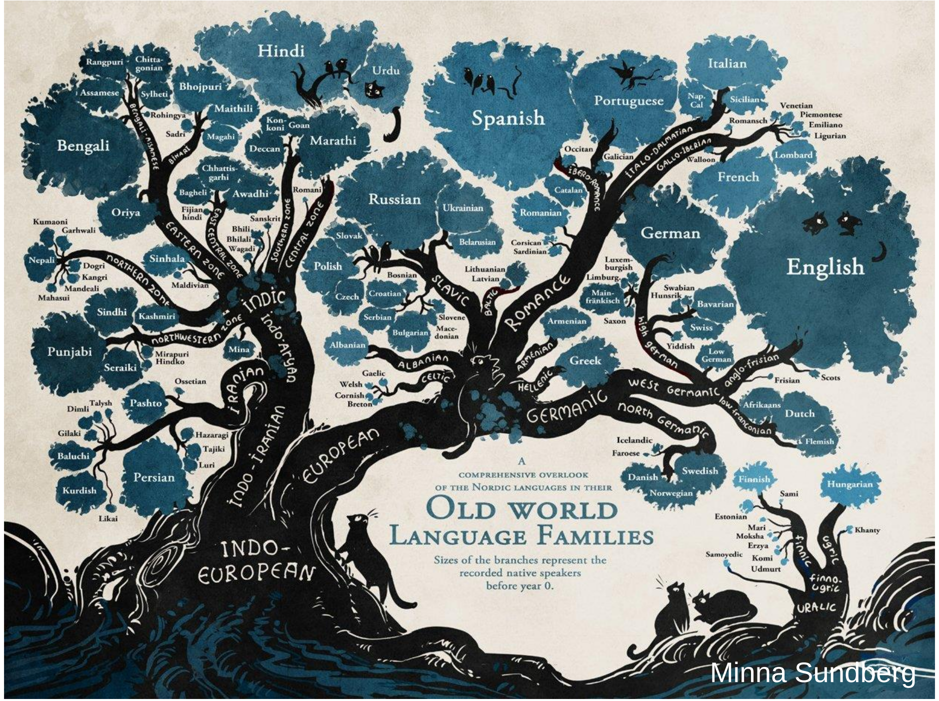
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
    - 12 fusional
    - 3 analytic
    - 1 isolating
    - 7 agglutinative



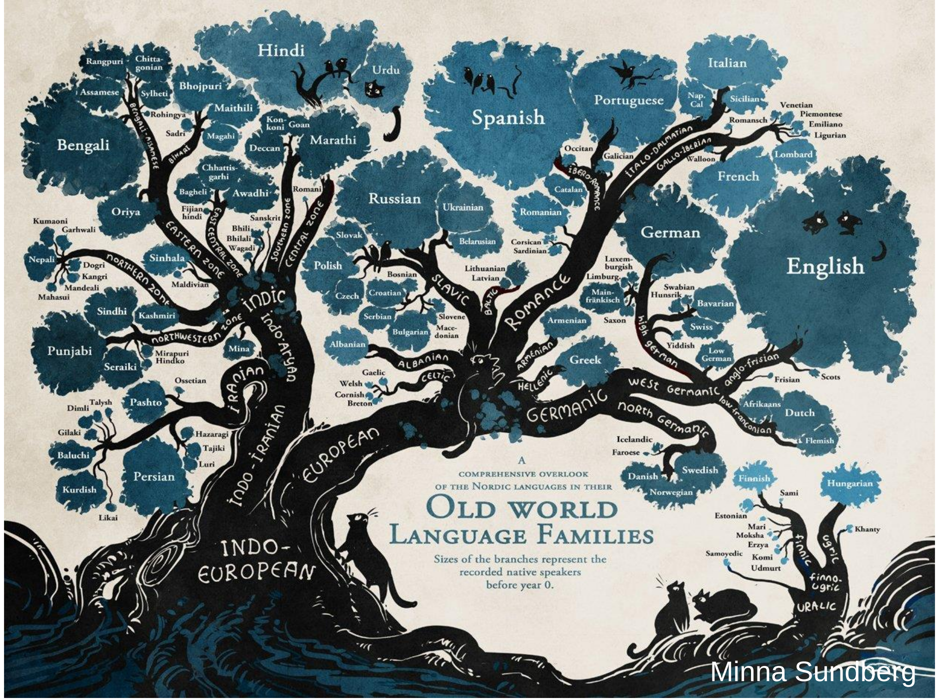
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity



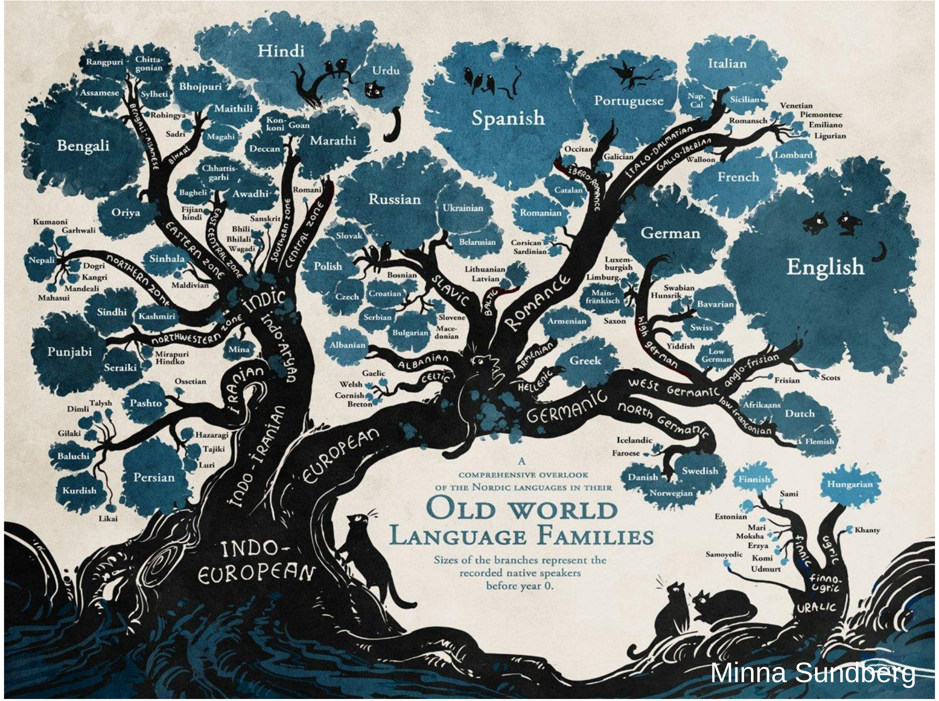
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)



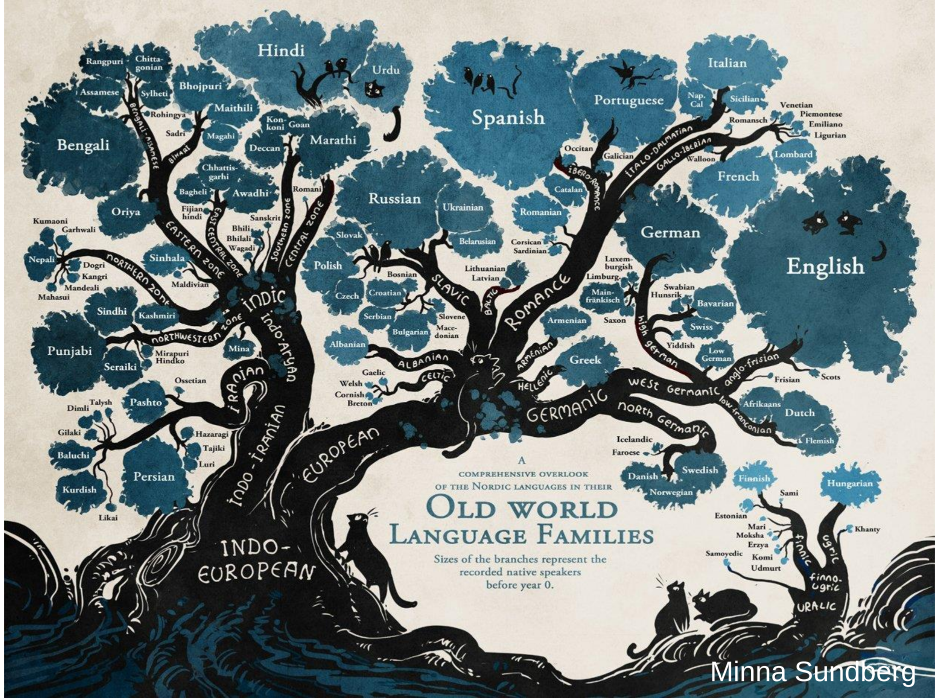
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)
  - 10 from 8 non-IE branches



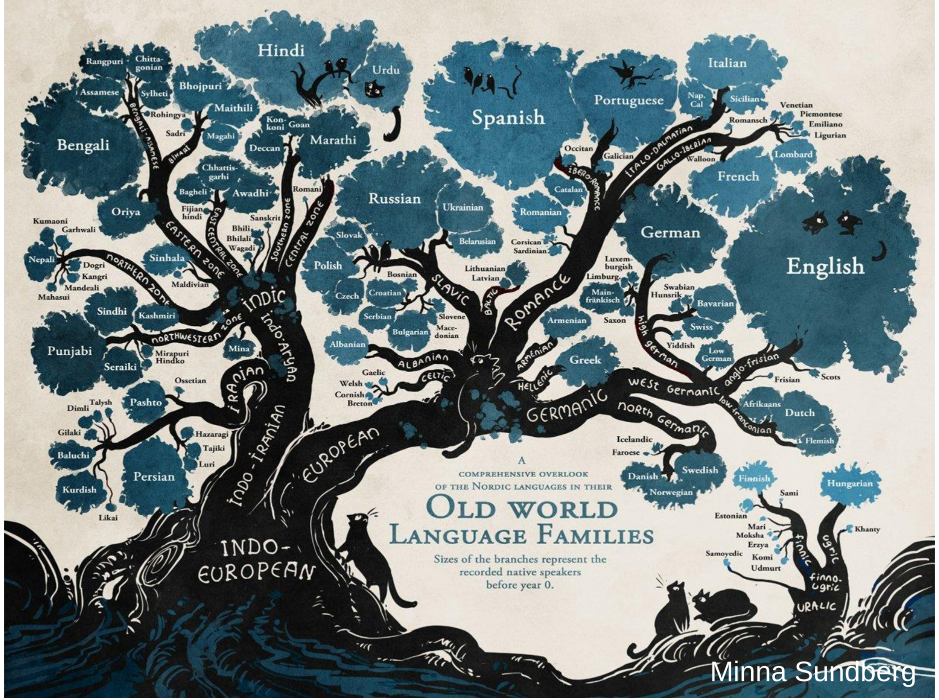
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)
  - 10 from 8 non-IE branches
- MRLs (e.g. Slavic languages)



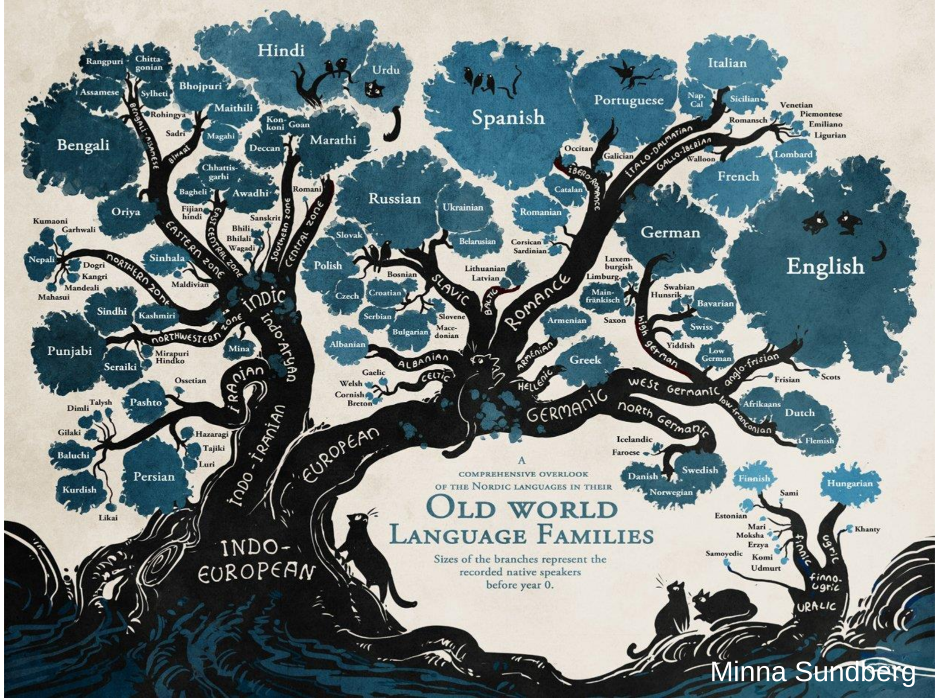
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)
  - 10 from 8 non-IE branches
- MRLs (e.g. Slavic languages)
  - Much word-level data



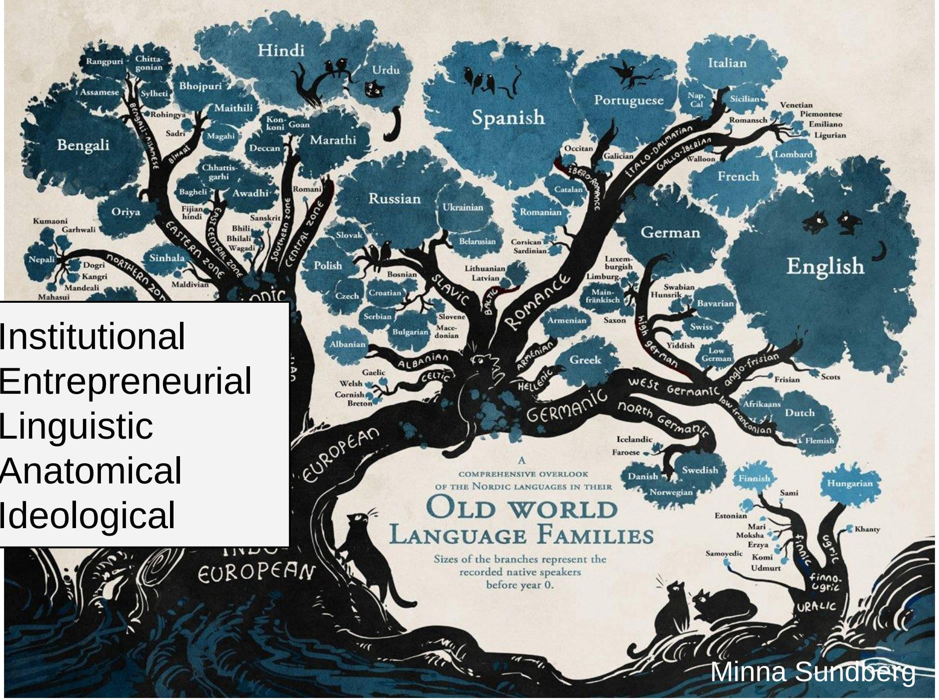
Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)
  - 10 from 8 non-IE branches
- MRLs (e.g. Slavic languages)
  - Much word-level data
  - Relatively free word order



Language Selection

- |UD ∩ Polyglot| = 44, we took 23
- Morphological structure
  - 12 fusional
  - 3 analytic
  - 1 isolating
  - 7 agglutinative
- Geneological diversity
  - 13 Indo-European (7 different branches)
  - 10 from 8 non-IE branches
- MRLs (e.g. Slavic languages)
  - Much word-level data
  - Relatively free word order



Institutional
Entrepreneurial
Linguistic
Anatomical
Ideological

Language Selection (contd.)

Language Selection (contd.)

- Script type

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables
- Attribute-carrying tokens

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables
- Attribute-carrying tokens
  - Range from 0% (Vietnamese) to 92.4% (Hindi)

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables
- Attribute-carrying tokens
  - Range from 0% (Vietnamese) to 92.4% (Hindi)
- OOV rate (UD against Polyglot vocabulary)

Language Selection (contd.)

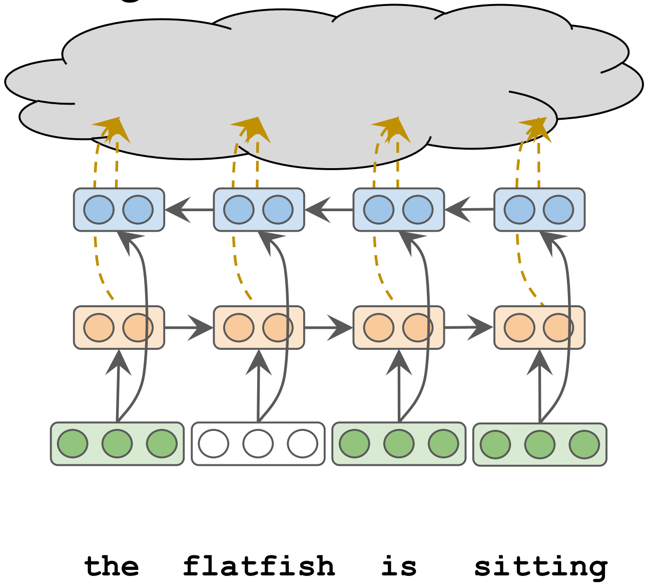
- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables
- Attribute-carrying tokens
  - Range from 0% (Vietnamese) to 92.4% (Hindi)
- OOV rate (UD against Polyglot vocabulary)
  - 16.9%-70.8% type-level (median 29.1%)

Language Selection (contd.)

- Script type
  - 7 in non-alphabetic scripts
  - Ideographic (Chinese) - ~12K characters
  - Hebrew, Arabic - no casing, no vowels, syntactic fusion
  - Vietnamese - tokens are non-compositional syllables
- Attribute-carrying tokens
  - Range from 0% (Vietnamese) to 92.4% (Hindi)
- OOV rate (UD against Polyglot vocabulary)
  - 16.9%-70.8% type-level (median 29.1%)
  - 2.2%-33.1% token-level (median 9.2%)

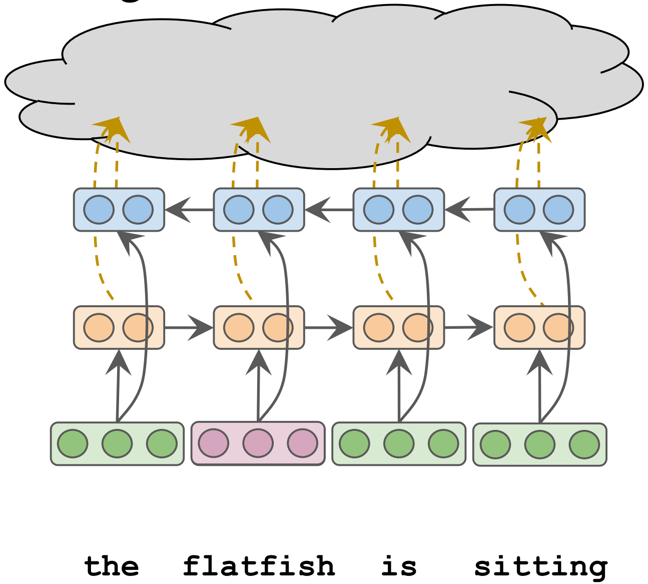
Evaluated Systems

- NONE: Polyglot's default UNK embedding



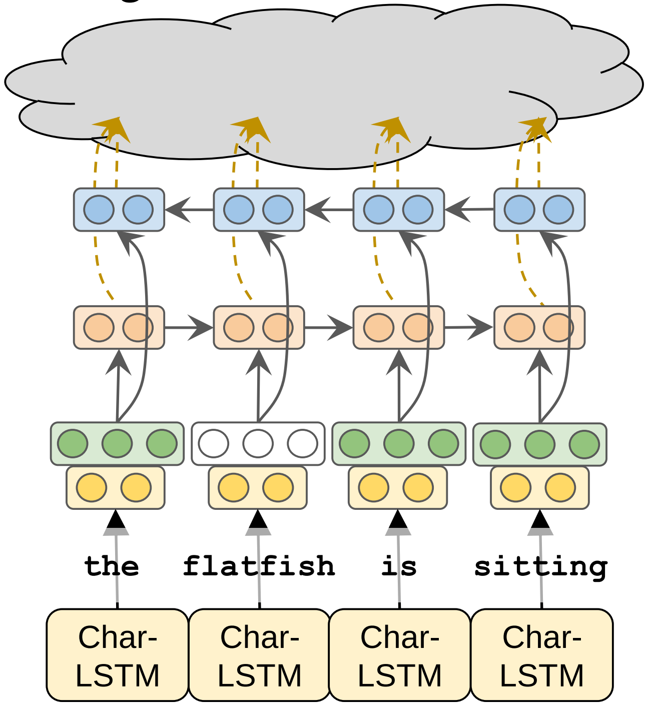
Evaluated Systems

- NONE: Polyglot's default UNK embedding
- MIMICK



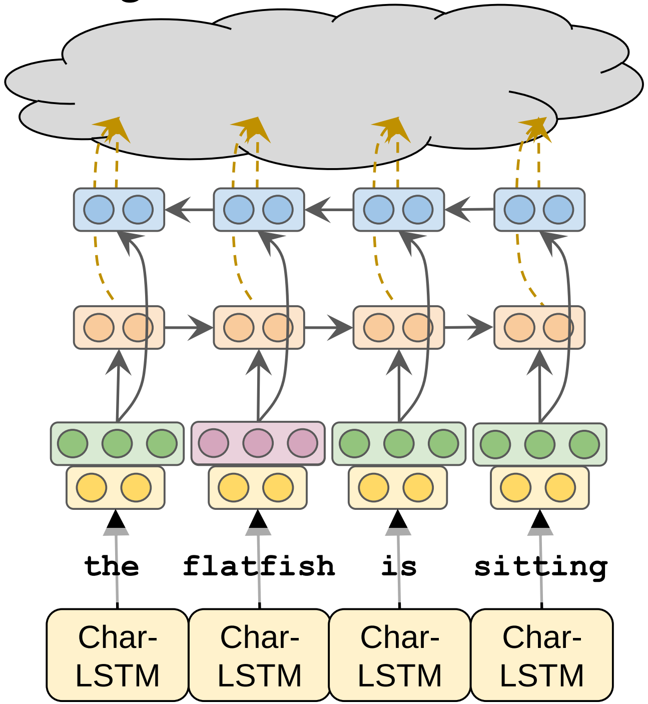
Evaluated Systems

- NONE: Polyglot's default UNK embedding
- MIMICK
- CHAR2TAG - additional RNN layer
  - 3x Training time



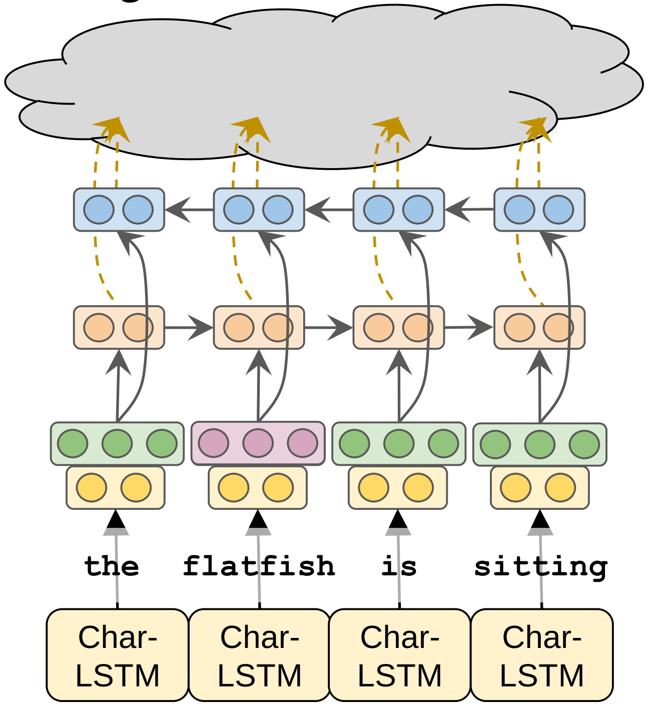
Evaluated Systems

- NONE: Polyglot's default UNK embedding
- MIMICK
- CHAR2TAG - additional RNN layer
  - 3x Training time
- BOTH: MIMICK + CHAR2TAG



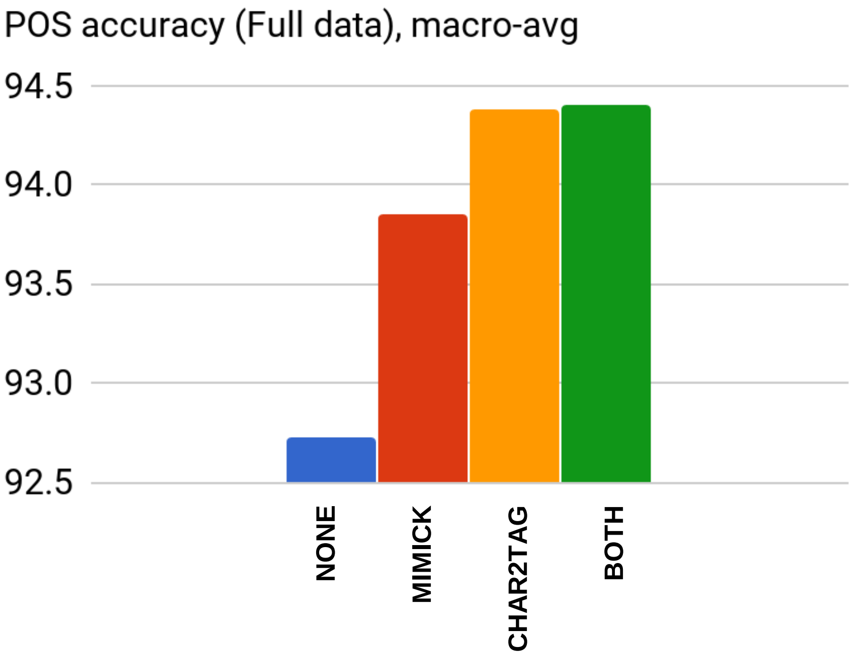
Evaluated Systems

- NONE: Polyglot's default UNK embedding
- MIMICK
- CHAR2TAG - additional RNN layer
  - 3x Training time
- BOTH: MIMICK + CHAR2TAG

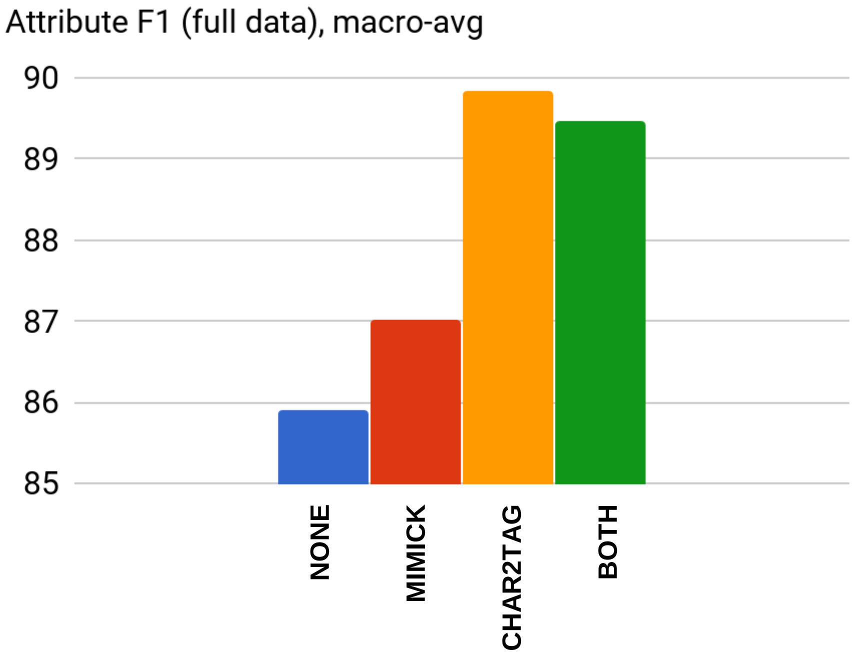


POINT
UNION
ROAD
LIGHT
LONG

Results - Full Data

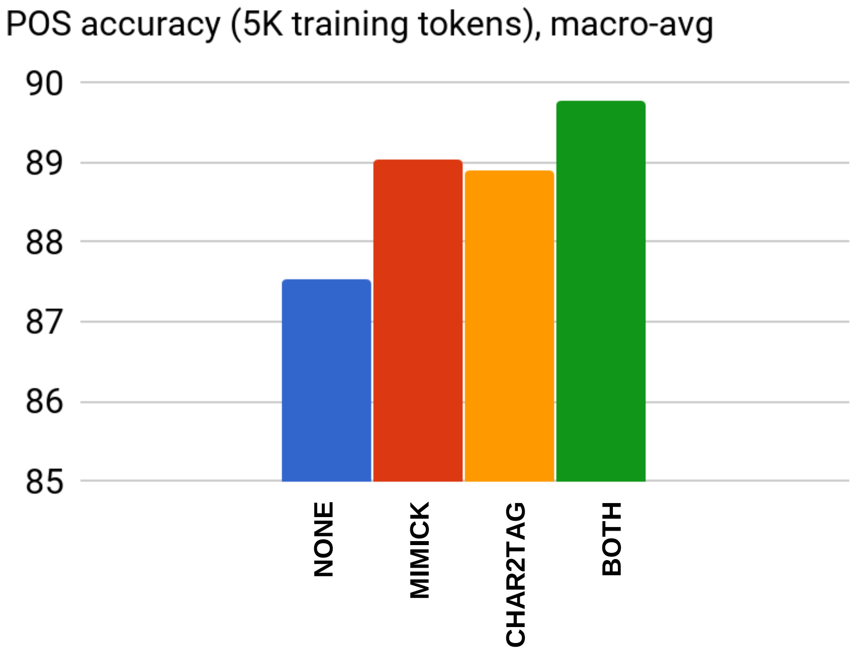


POS tags (accuracy)

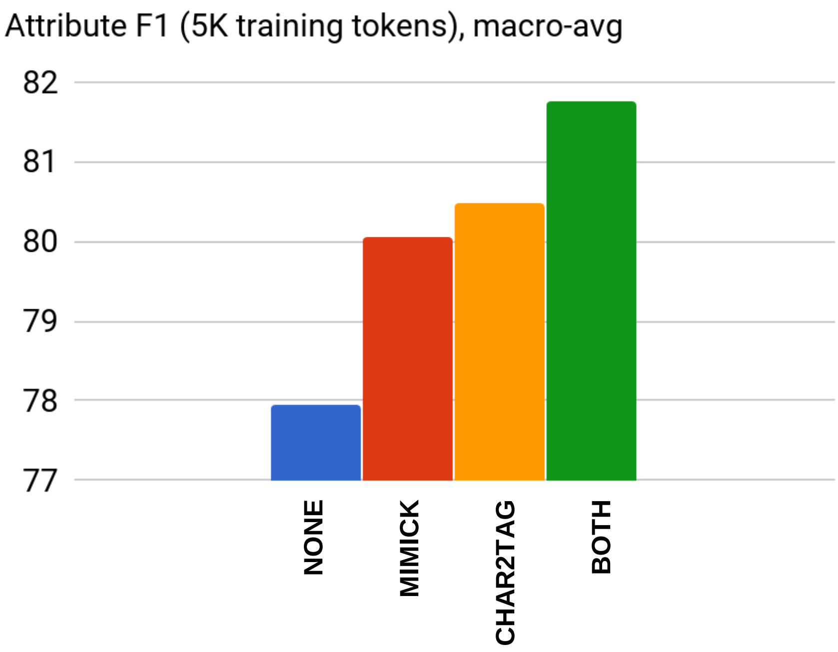


Morpho. Attributes (micro F1)

Results - 5,000 training tokens

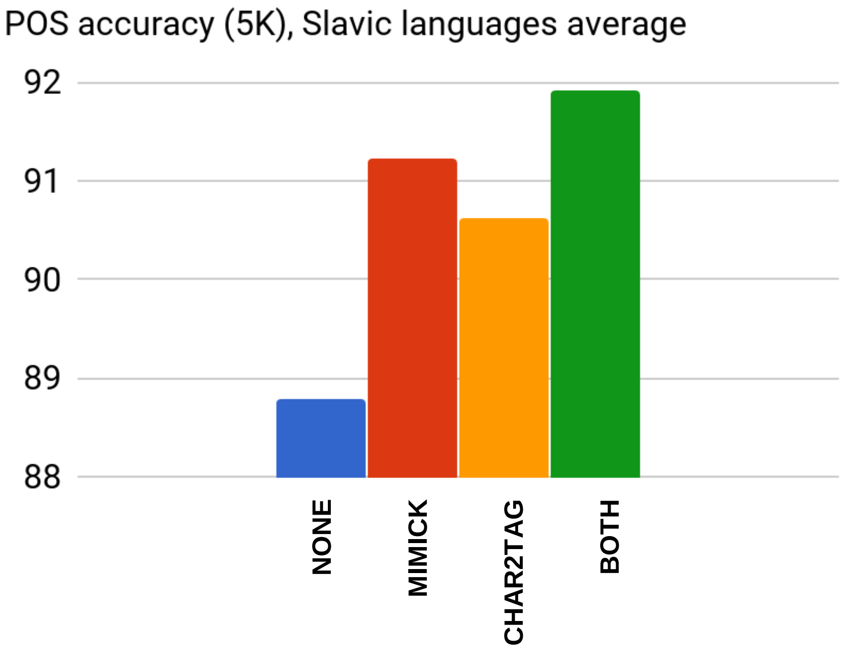


POS tags (accuracy)



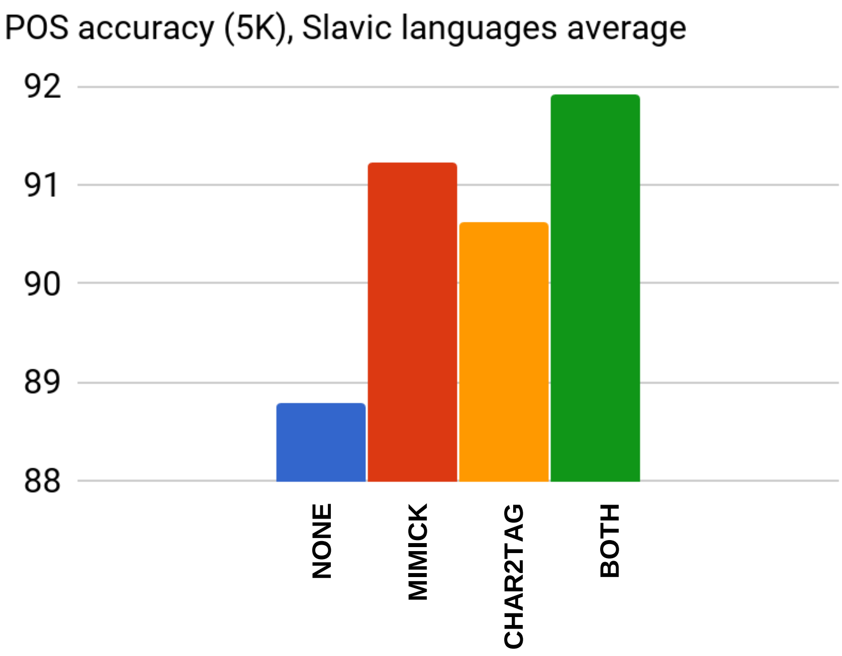
Morpho. Attributes (micro F1)

Results - Language Types (5,000 tokens)

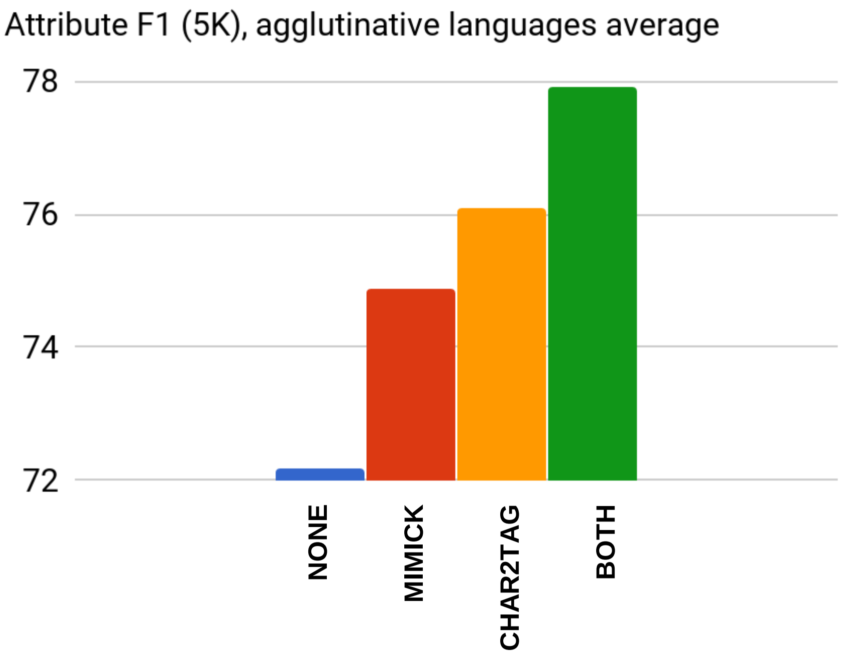


Slavic languages POS

Results - Language Types (5,000 tokens)

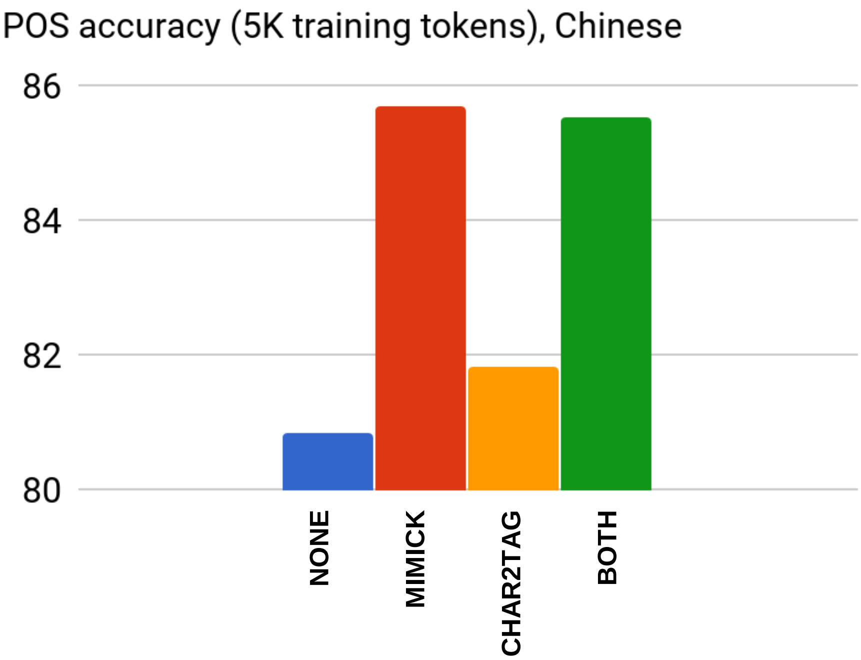


Slavic languages POS

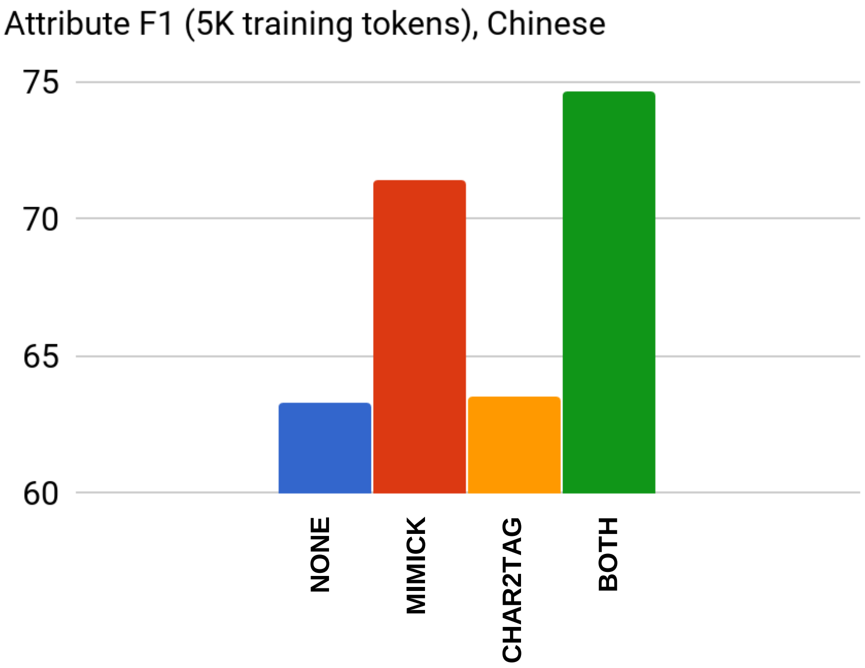


Agglutinative languages morpho. attribute F1

Results - Chinese



POS tags (accuracy)



Morpho. Attributes (micro F1)

A Word (Model) from our Sponsor

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
    - Sentiment!
    - Parsing!
    - IE!

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!
  - IE!
  - QA!

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
    - Sentiment!
    - Parsing!
    - IE!
    - QA!
    - …

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!
  - IE!
  - QA!
  - …
- Code compatible with w2v, Polyglot, FastText

Code & models:
https://github.com/yuvalpinter/Mimick

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!
  - IE!
  - QA!
  - ...
- Code compatible with w2v, Polyglot, FastText
- Models for Polyglot also on github

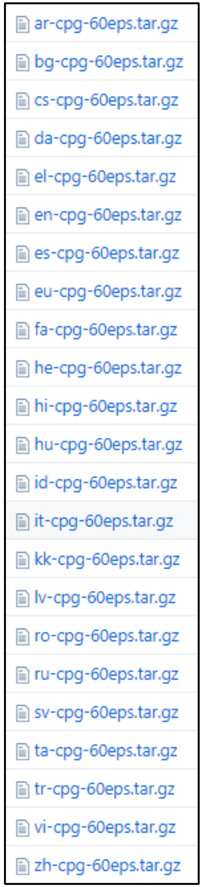
Code & models:
https://github.com/yuvalpinter/Mimick

- ar-cpg-60eps.tar.gz
- bg-cpg-60eps.tar.gz
- cs-cpg-60eps.tar.gz
- da-cpg-60eps.tar.gz
- el-cpg-60eps.tar.gz
- en-cpg-60eps.tar.gz
- es-cpg-60eps.tar.gz
- eu-cpg-60eps.tar.gz
- fa-cpg-60eps.tar.gz
- he-cpg-60eps.tar.gz
- hi-cpg-60eps.tar.gz
- hu-cpg-60eps.tar.gz
- id-cpg-60eps.tar.gz
- it-cpg-60eps.tar.gz
- kk-cpg-60eps.tar.gz
- lv-cpg-60eps.tar.gz
- ro-cpg-60eps.tar.gz
- ru-cpg-60eps.tar.gz
- sv-cpg-60eps.tar.gz
- ta-cpg-60eps.tar.gz
- tr-cpg-60eps.tar.gz
- vi-cpg-60eps.tar.gz
- zh-cpg-60eps.tar.gz

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
    - Sentiment!
    - Parsing!
    - IE!
    - QA!
    - ...
- Code compatible with w2v, Polyglot, FastText
- Models for Polyglot also on github
    - <1MB each, dynet format

Code & models:
https://github.com/yuvalpinter/Mimick



A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!
  - IE!
  - QA!
  - ...
- Code compatible with w2v, Polyglot, FastText
- Models for Polyglot also on github
  - <1MB each, dynet format
  - Learn all OOVs in advance and add to param table, **or**

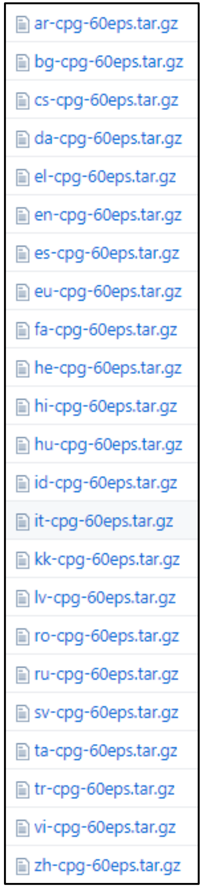
Code & models:
https://github.com/yuvalpinter/Mimick

- ar-cpg-60eps.tar.gz
- bg-cpg-60eps.tar.gz
- cs-cpg-60eps.tar.gz
- da-cpg-60eps.tar.gz
- el-cpg-60eps.tar.gz
- en-cpg-60eps.tar.gz
- es-cpg-60eps.tar.gz
- eu-cpg-60eps.tar.gz
- fa-cpg-60eps.tar.gz
- he-cpg-60eps.tar.gz
- hi-cpg-60eps.tar.gz
- hu-cpg-60eps.tar.gz
- id-cpg-60eps.tar.gz
- it-cpg-60eps.tar.gz
- kk-cpg-60eps.tar.gz
- lv-cpg-60eps.tar.gz
- ro-cpg-60eps.tar.gz
- ru-cpg-60eps.tar.gz
- sv-cpg-60eps.tar.gz
- ta-cpg-60eps.tar.gz
- tr-cpg-60eps.tar.gz
- vi-cpg-60eps.tar.gz
- zh-cpg-60eps.tar.gz

A Word (Model) from our Sponsor

- Our extrinsic results are on **tagging**
- Please consider us for all your WE use cases!
  - Sentiment!
  - Parsing!
  - IE!
  - QA!
  - ...
- Code compatible with w2v, Polyglot, FastText
- Models for Polyglot also on github
  - <1MB each, dynet format
  - Learn all OOVs in advance and add to param table, **or**
  - Load into memory and infer on-line

Code & models:
https://github.com/yuvalpinter/Mimick



Conclusions

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
    - Morphologically-rich languages

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
    - Morphologically-rich languages
    - Large character vocabulary

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
    - Morphologically-rich languages
    - Large character vocabulary
- Sore spots and Future Work

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
    - Morphologically-rich languages
    - Large character vocabulary
- Sore spots and Future Work
    - Vietnamese - syllabic vocabulary

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
  - Morphologically-rich languages
  - Large character vocabulary
- Sore spots and Future Work
  - Vietnamese - syllabic vocabulary
  - Hebrew and Arabic - nontrivial tokenization, no case

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
  - Morphologically-rich languages
  - Large character vocabulary
- Sore spots and Future Work
  - Vietnamese - syllabic vocabulary
  - Hebrew and Arabic - nontrivial tokenization, no case
  - Try other subword levels (morphemes, phonemes, bytes)

Conclusions

- MIMICK: an OOV-extension embedding processing step for downstream tasks
- Compositional model complementing distributional artifact
- Powerful technique for **low-resource** scenarios
- Especially good for:
  - Morphologically-rich languages
  - Large character vocabulary
- Sore spots and Future Work
  - Vietnamese - syllabic vocabulary
  - Hebrew and Arabic - nontrivial tokenization, no case
  - Try other subword levels (morphemes, phonemes, bytes)
  - Improve morphosyntactic attribute tagging scheme

Questions?

Neglect
Satisfaction
Illness
Espionage
Bullying

Code & models:
https://github.com/yuvalpinter/Mimick