Statistical Machine Translation LING-462/COSC-482 Week 11: Domain Adaptation and Word Embedding

Achim Ruopp
achim.ruopp@Georgetown.edu

Agenda

- Language in ten minutes: American Sign Language – Emma Manning
- Domain Adaptation

 - Break -

- Word Embedding and Word2Vec
- Homework 5 Suggestions

”Domain”

- Corpora differ
  - topic (politics, news, medicine, ...)
  - style (formal, informal)
  - modality (written, transcribed speech)
  - register (level of politeness)

- Covered on the catch-all term ”domain”

- Domain := one source for a parallel corpus

"Domain"

- Domain matters for word choice
  - *bat* in baseball domain vs. *bat* in animal domain
  - *interest* in financial domain vs. *interest* in arts

- Style matters, too
  - translate greeting into *What's up?* vs. *Ladies and Gentlemen!*
  - use of informal *Du* vs. formal *Sie* in German

- Distinctions often only visible in full document / full corpus

Various Data Sources

- Available parallel corpora on OPUS web site (Italian–English)

corpus	doc's	sent's	it tokens	en tokens	XCES/XML	raw	TMX	Moses
OpenSubtitles2018	48746	37.8M	304.8M	284.5M	[xces en it]	[en it]	[tmx]	[moses]
EUbookshop	9028	6.6M	268.7M	258.8M	[xces en it]	[en it]	[tmx]	[moses]
OpenSubtitles2016	35929	28.7M	230.3M	214.9M	[xces en it]	[en it]	[tmx]	[moses]
DGT	26880	3.2M	72.9M	64.0M	[xces en it]	[en it]	[tmx]	[moses]
Europarl	9461	2.0M	59.9M	58.9M	[xces en it]	[en it]	[tmx]	[moses]
JRC-Acquis	12042	0.8M	34.1M	34.5M	[xces en it]	[en it]	[tmx]	[moses]
Wikipedia	3	1.0M	26.5M	22.2M	[xces en it]	[en it]	[tmx]	[moses]
EMEA	1920	1.1M	12.0M	13.9M	[xces en it]	[en it]	[tmx]	[moses]
ECB	1	0.2M	5.5M	5.8M	[xces en it]	[en it]	[tmx]	[moses]
GNOME	1905	0.7M	3.8M	3.4M	[xces en it]	[en it]	[tmx]	[moses]
TED2013	1	0.2M	3.2M	2.7M	[xces en it]	[en it]	[tmx]	[moses]
Tanzil	15	0.1M	2.8M	2.4M	[xces en it]	[en it]	[tmx]	[moses]
Tatoeba	1	0.1M	3.6M	1.3M	[xces en it]	[en it]	[tmx]	[moses]
KDE4	1957	0.3M	2.2M	2.3M	[xces en it]	[en it]	[tmx]	[moses]
GlobalVoices	3220	81.3k	2.1M	2.0M	[xces en it]	[en it]	[tmx]	[moses]
News-Commentary11	1423	45.9k	1.3M	1.0M	[xces en it]	[en it]	[tmx]	[moses]
Books	8	33.1k	0.9M	0.8M	[xces en it]	[en it]	[tmx]	[moses]
Ubuntu	452	0.1M	0.8M	0.6M	[xces en it]	[en it]	[tmx]	[moses]
News-Commentary	1	18.6k	0.5M	0.5M	[xces en it]	[en it]	[tmx]	[moses]
PHP	3270	36.8k	0.5M	0.2M	[xces en it]	[en it]	[tmx]	[moses]
EUconst	47	10.2k	0.2M	0.2M	[xces en it]	[en it]	[tmx]	[moses]
OpenSubtitles	22	19.1k	0.2M	0.1M	[xces en it]	[en it]	[tmx]	[moses]
total	**156332**	**83.1M**	**1.0G**	**975.1M**	**83.1M**		**63.4M**	**77.4M**

Domain Examples

EMEA Abilify is a medicine containing the active substance aripiprazole.
It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets that dissolve in the mouth), as an oral solution (1 mg/ml) and as a solution for injection (7.5 mg/ml).

Software Localization Default GNOME Theme
OK
People
Pictures
Plan
Sound

Literature There was a slight noise behind her and she turned just in time to seize a small boy by the slack of his roundabout and arrest his flight.

Law Corrigendum to the Interim Agreement with a view to an Economic Partnership Agreement between the European Community and its Member States, of the one part, and the Central Africa Party, of the other part.

Domain Examples

PHP If you would like to start a new translation, or help in a translation project, please read http://cvs.php.net/co.php/phpdoc/howto/howto.html.tar.gz.

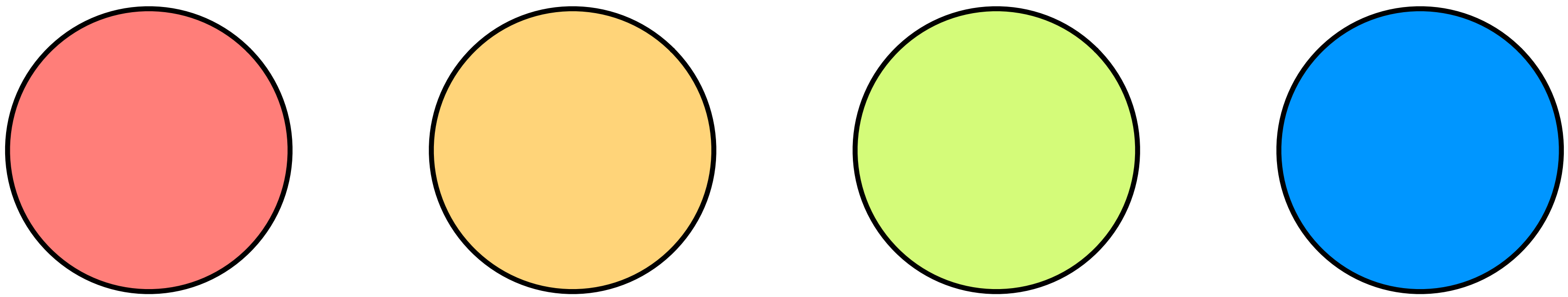
Religion This is The Book free of doubt and involution, a guidance for those who preserve themselves from evil and follow the straight path.

News The Facebook page of a leading Iranian leading cartoonist, Mana Nayestani, was hacked on Tuesday, 11 September 2012, by pro-regime hackers who call themselves "Soldiers of Islam".

Movie subtitles We're taking you to Washington, D.C.
Do you know where the prisoner was transported to?
Uh, Washington.
Okay.

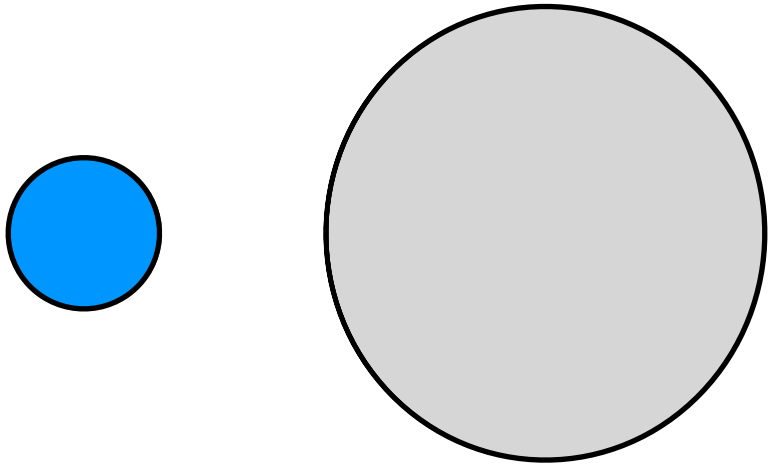
Twitter Thank u @Starbucks & @Spotify for celebrating artists who #GiveGood with a donation to @BTWFoundation, and to great organizations by @Metallica and @ChanceTheRapper! Limited edition cards available now at Starbucks!

Multi-Domain Scenario



- Machine translation systems work best when optimized for one domain
- Separate data by domain
- Build special system for each domain
- Translate each sentence with matching system

In/Out Domain Scenario



- Optimize system for just one domain
- Available data
  - small amounts of in-domain data
  - large amounts of out-of-domain data
- Need to balance both data sources

Why Use Out-of-Domain Data?

- In-domain data much more valuable
- But: gaps
  - word-to-be-translated may not occur
  - word-to-be-translated may not occur with the correct translation
- Motivation
  - out-of-domain data may fill these gaps
  - but be careful not to drown out in-domain data

S^4 Taxonomy of Adaptation Effects

[Carpuat, Daume, Fraser, Quirk, 2012]

- **Seen**: Never seen this word before

  News to medical: diabetes mellitus

- **Sense**: Never seen this word used in this way

  News to technical: monitor

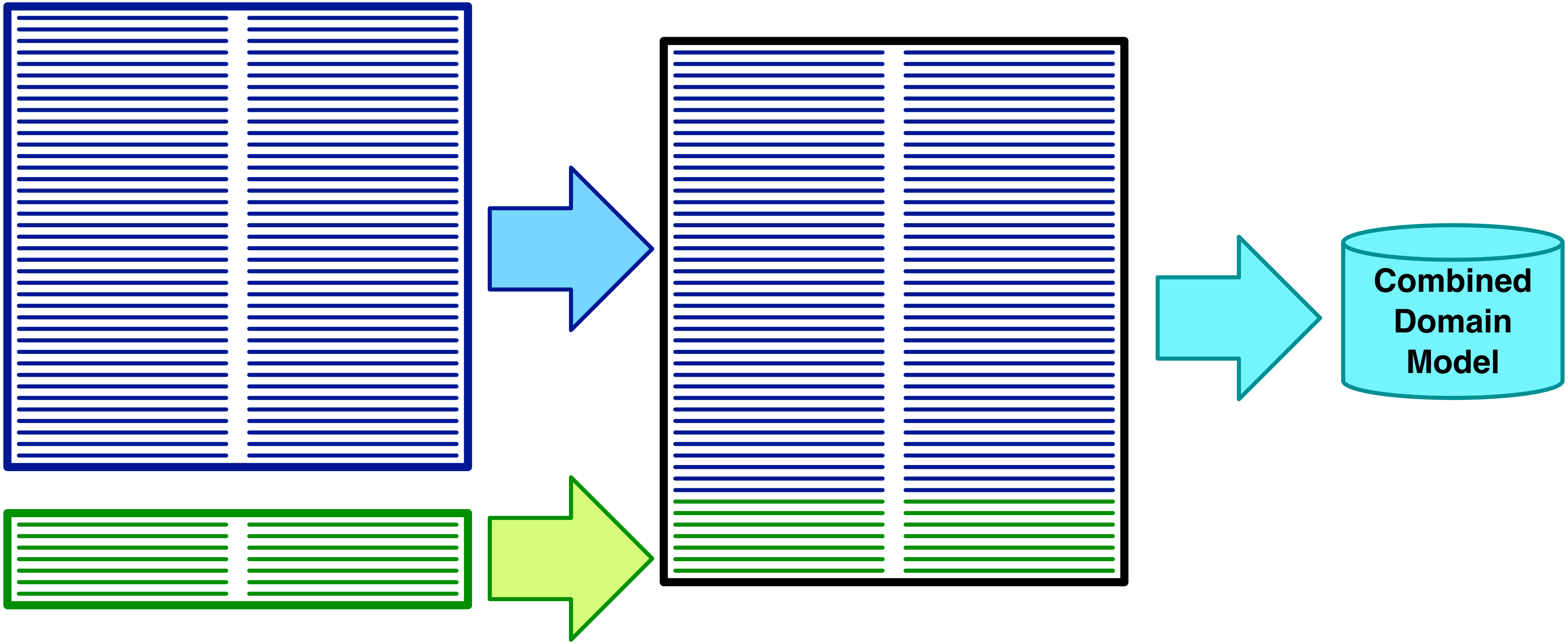
- **Score**: The wrong output is scored higher

  News to medical: manifest

- **Search**: Decoding/search erred

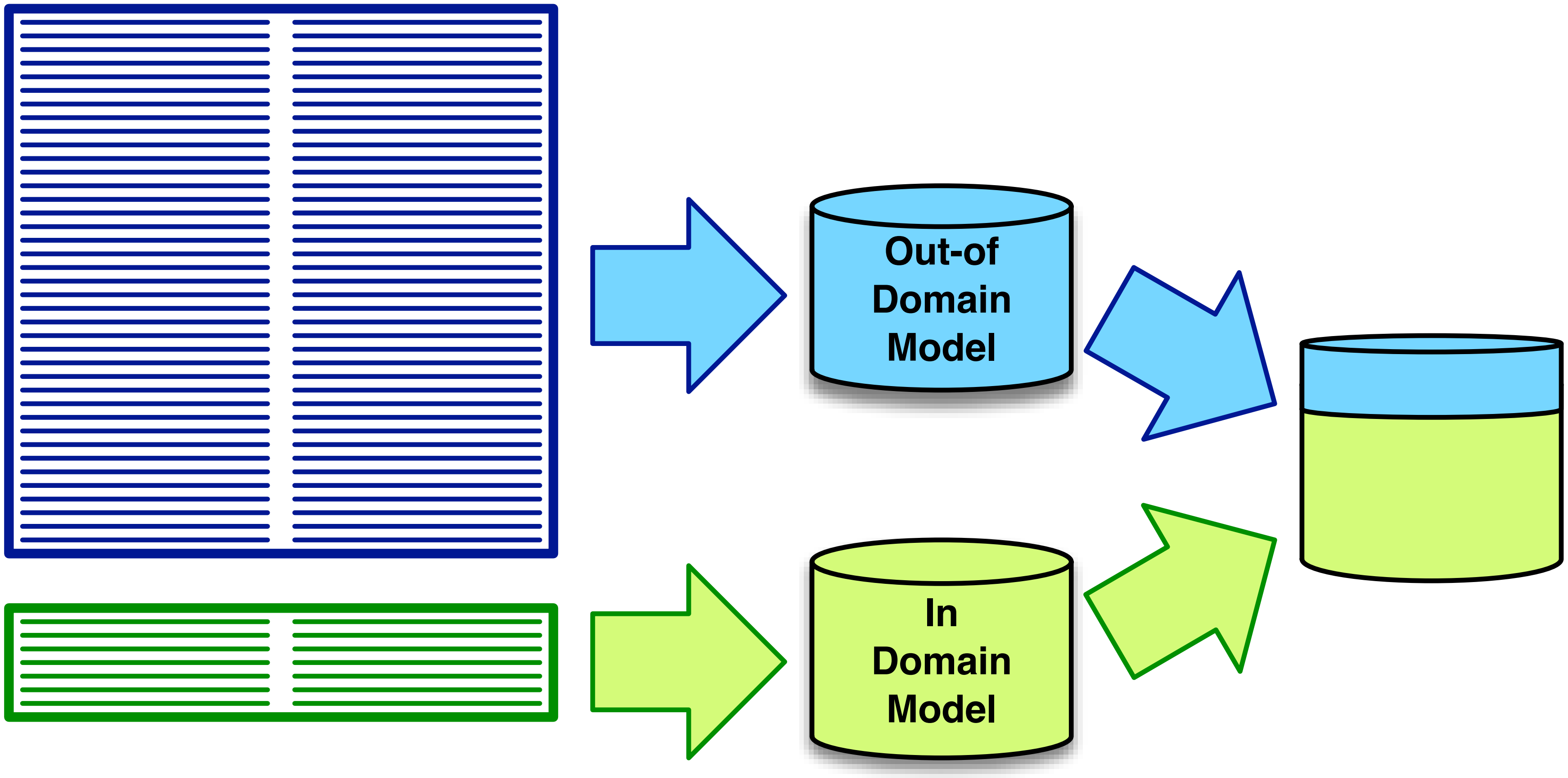
mixture models

Combining Data



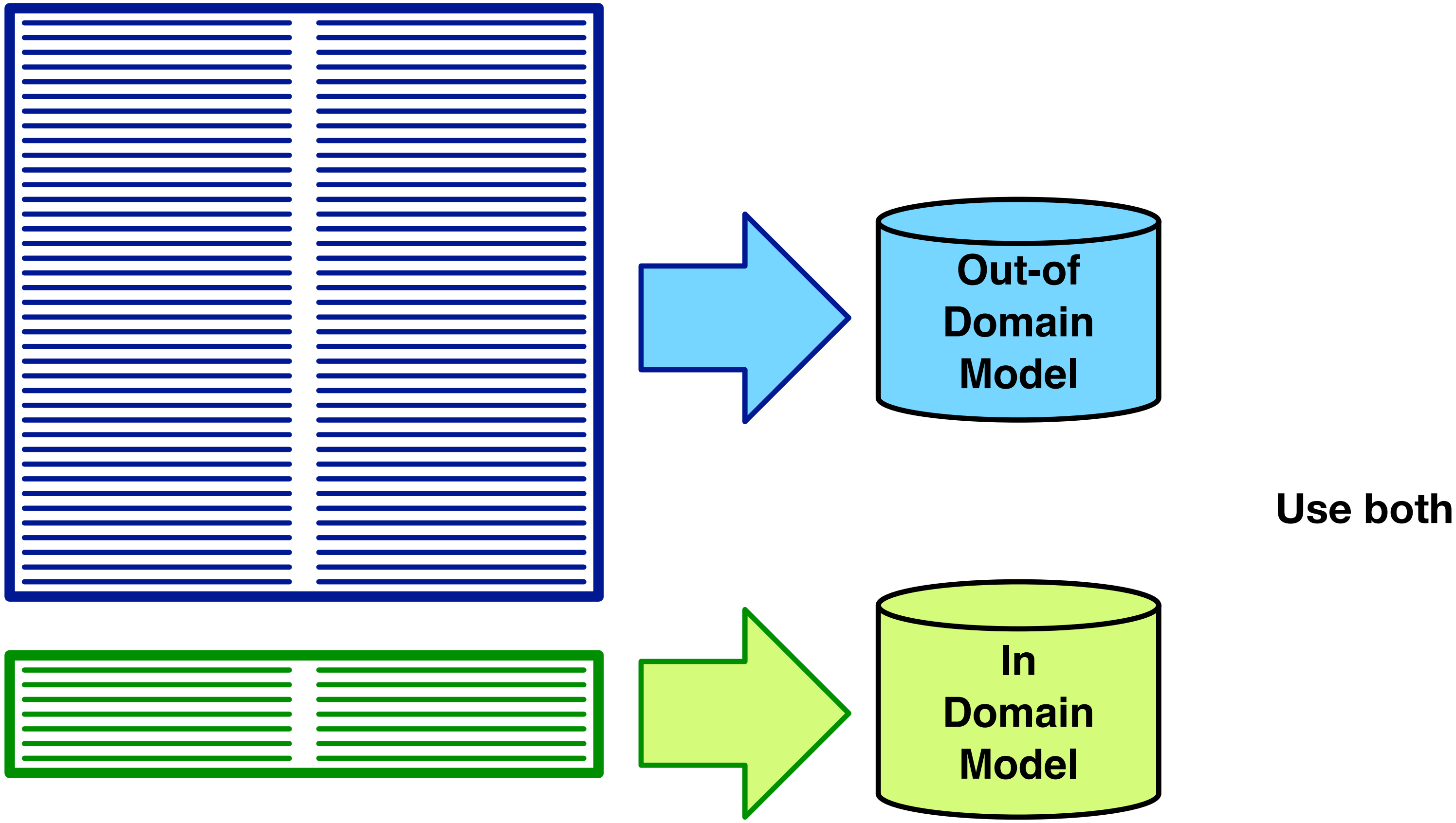
- Too biased towards out of domain data
- May flag translation options with indicator feature functions

Interpolate Models



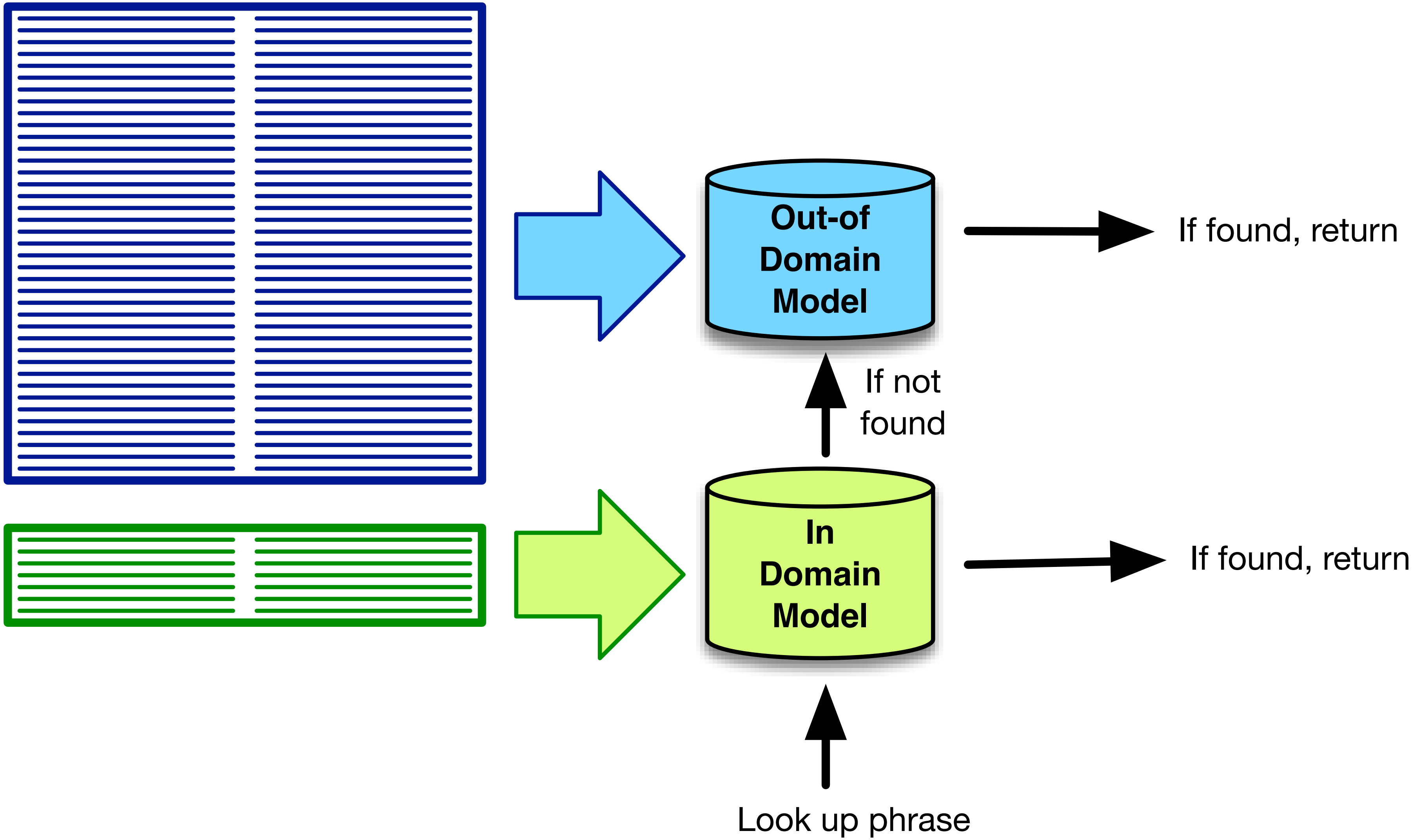
- $p_c(e|f) = \lambda_{\text{in}} p_{\text{in}}(e|f) + \lambda_{\text{out}} p_{\text{out}}(e|f)$
- Quite successful for language modelling

Multiple Models

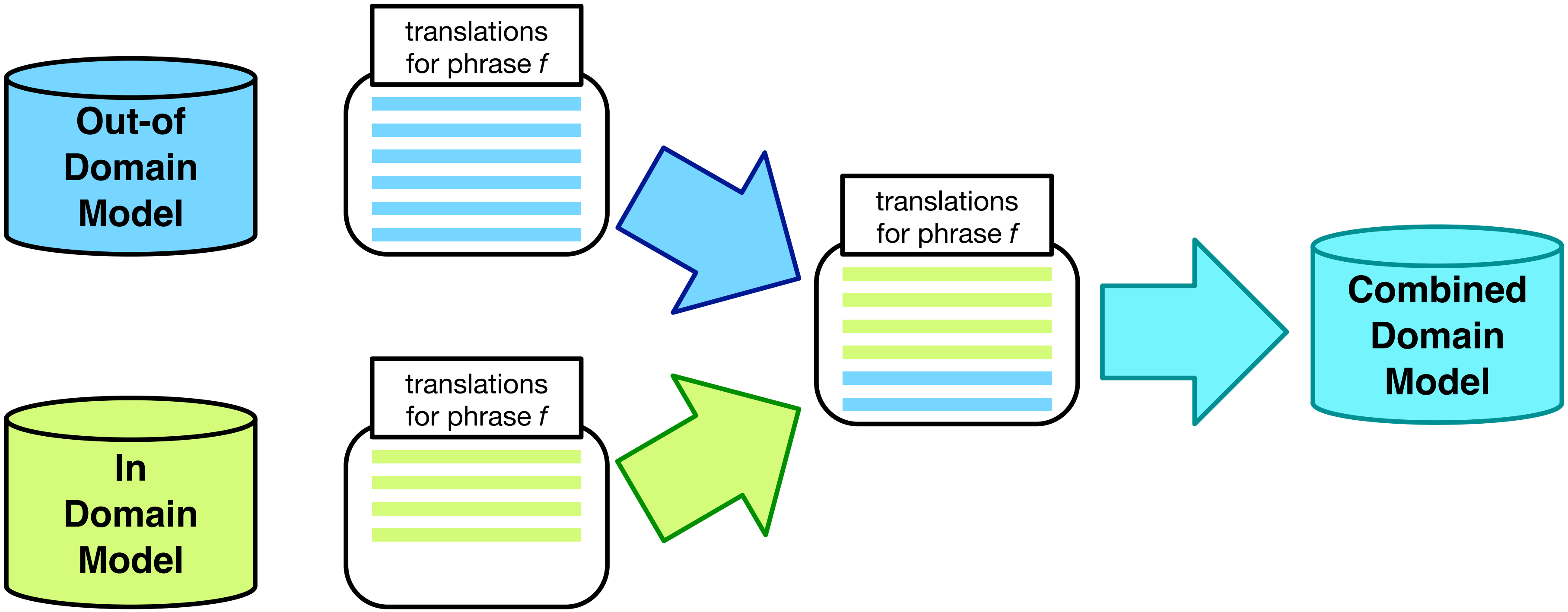


- Multiple models → multiple feature functions

Backoff

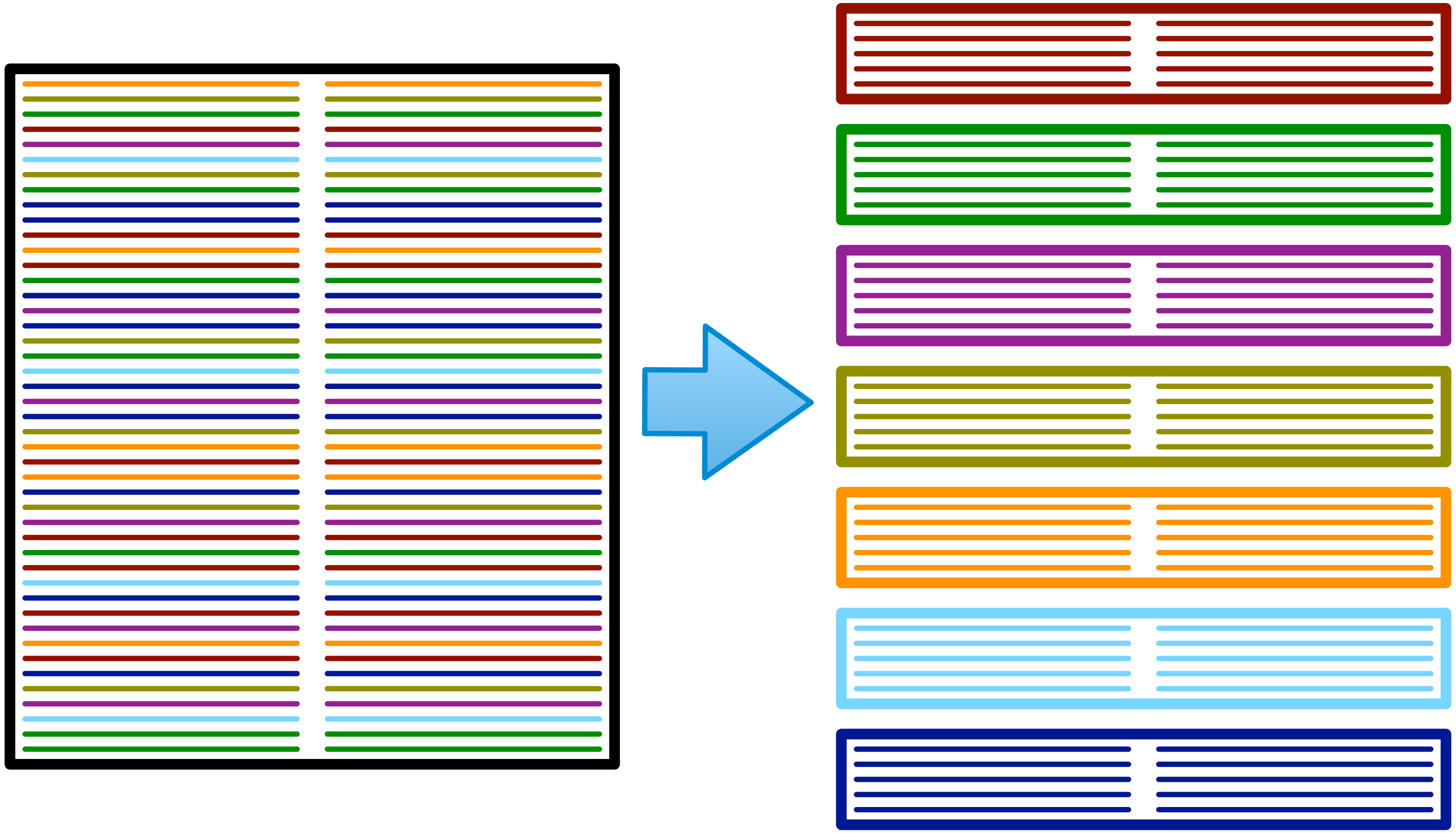

Out-of
Domain
Model
If found, return
If not
found
In
Domain
Model
If found, return
Look up phrase

Fill-Up



- Use translation options from in-domain table
- Fill up with additional options from out-of-domain table

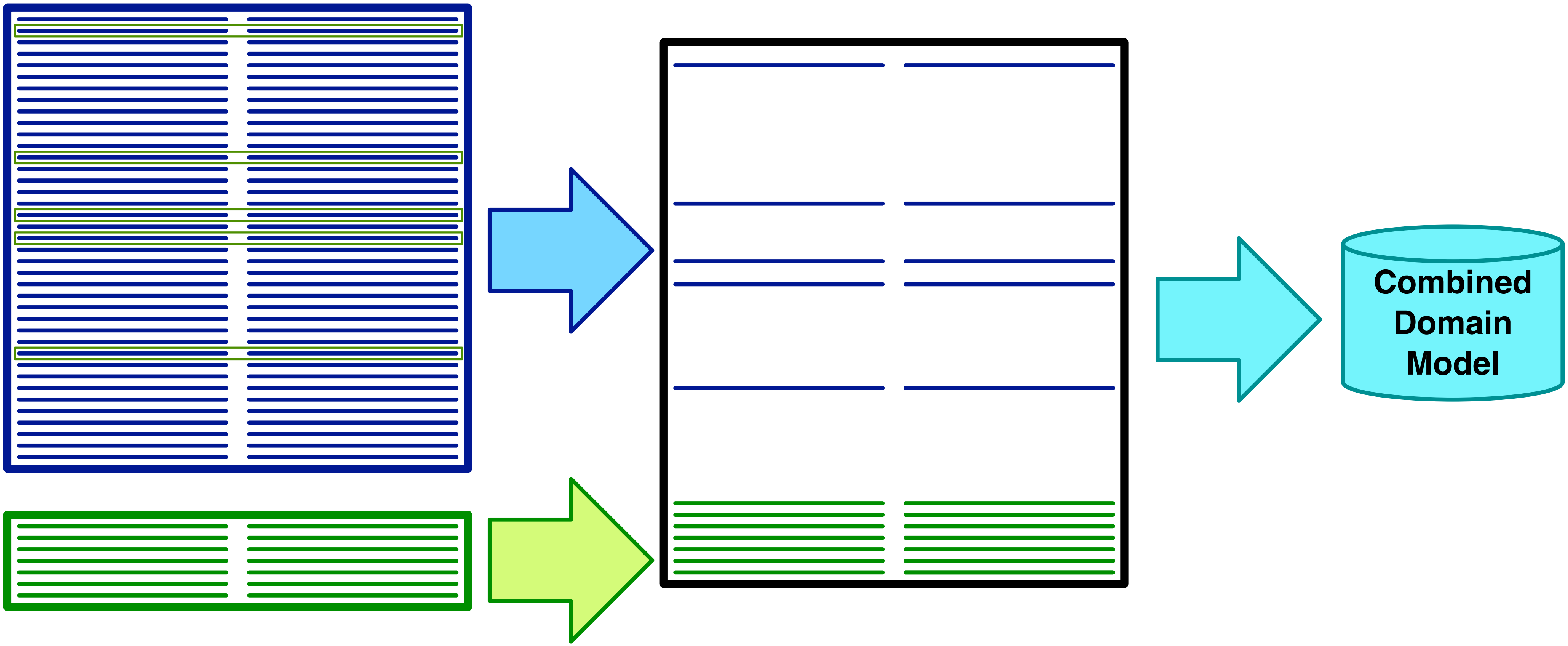
Topic Models



- Cluster corpus by topic — Latent Dirichlet Allocation (LDA)
- Train separate sub-models for each topic
- For input sentence, detect topic (or topic distribution)

subsampling

Sentence Selection

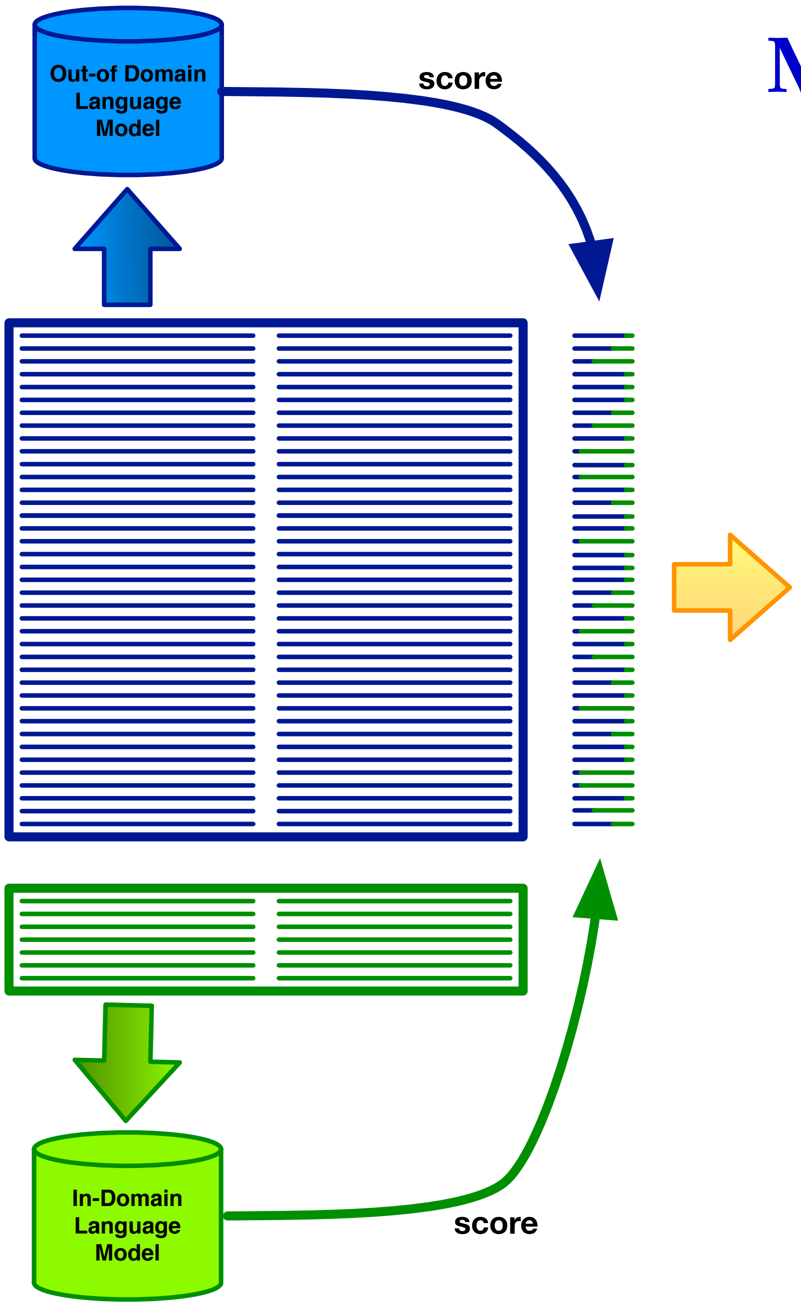


- Select out-of-domain sentence pairs that are similar to in-domain data

Sentence Selection

- Various methods
- Goal 1: Increase coverage (fill gaps)
- Goal 2: Get content with in-domain content, style, etc.

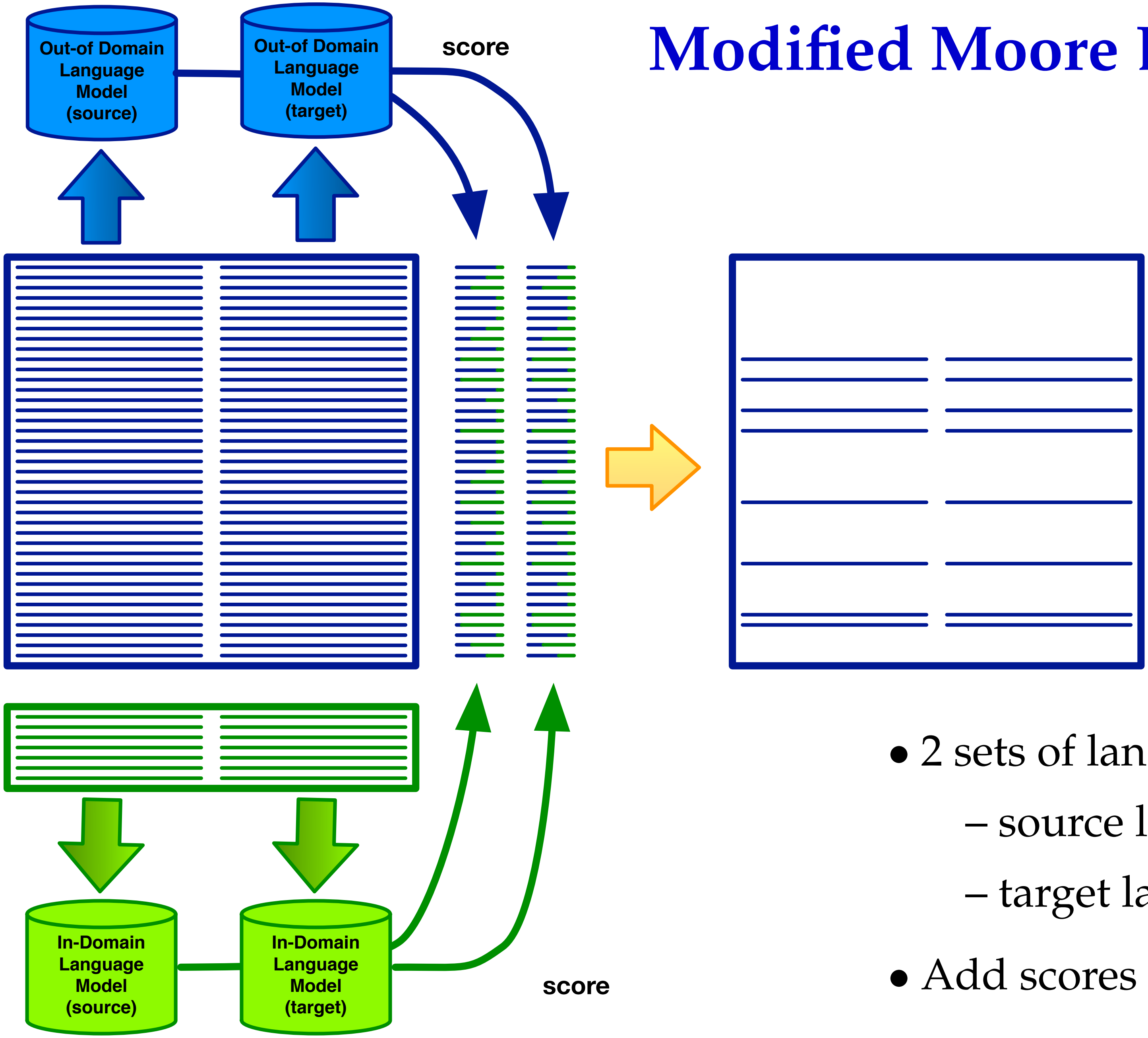
Moore Lewis



- Build language models
  - out of domain
  - in domain
- Score each sentence
- Sub-select sentence pairs with

$$p_{\text{IN}}(f) - p_{\text{OUT}}(f) > \tau$$

Modified Moore Lewis



- 2 sets of language models
  - source language
  - target language
- Add scores

Subsampling with POS

- Replace rare words with part-of-speech tags

an earthquake in Port-au-Prince

$\Downarrow$

an earthquake in NNP

- Works better [Axelrod et al., WMT2015]

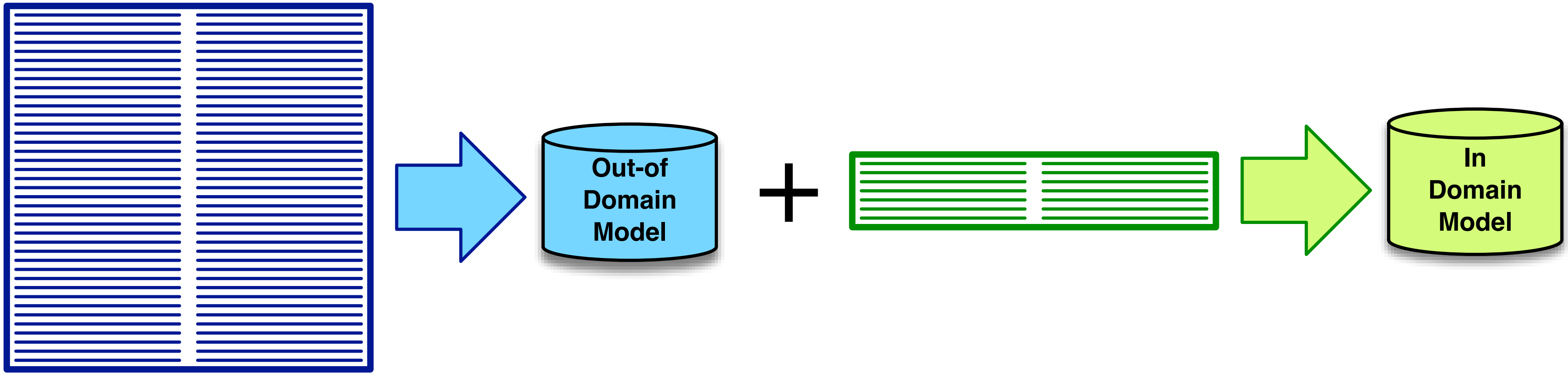
- Is it all about style, not key terminology?

Still Hard Problems

- How related are domains?
- Is corpus X useful for my system?
- What text properties matter?

neural adaptation

Fine Tuning



- First train system on out-of-domain data (or: all available data)
- Stop at convergence
- Then, continue training on in-domain data

- Successful even for fine tuning on 1 sentence pair [Farajian et al., WMT 2017]

Multi-Domain System

- Given: sets of corpora with known domain
- Task: translate sentence of known domain
  - training: add domain token, say [SPORTS], to each source sentence
  - testing: add domain token to input
- Task: translate sentence of unknown domain
  - training: learn separate models for each domain
  - testing: predict domain of sentence, weight ensemble of domain-models

Corpus Weighting

- Goal: Give more weight to in-domain data
- Solution: Duplicate in-domain data n times when merging
- But duplication factor not clear

Instance Weighting

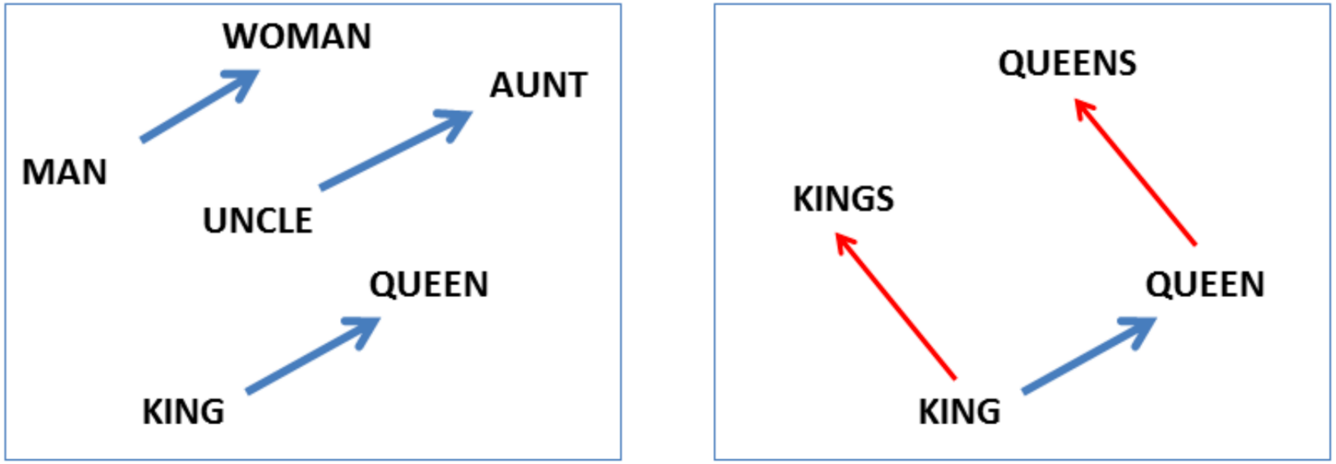
- For each sentence pair, compute domain-relatedness score (0–1)
  — could use something like Modified Moore-Lewis

- During training: scale learning rate based on this number

[Chen et al., NMT 2017]

WORD EMBEDDING AND WORD2VEC

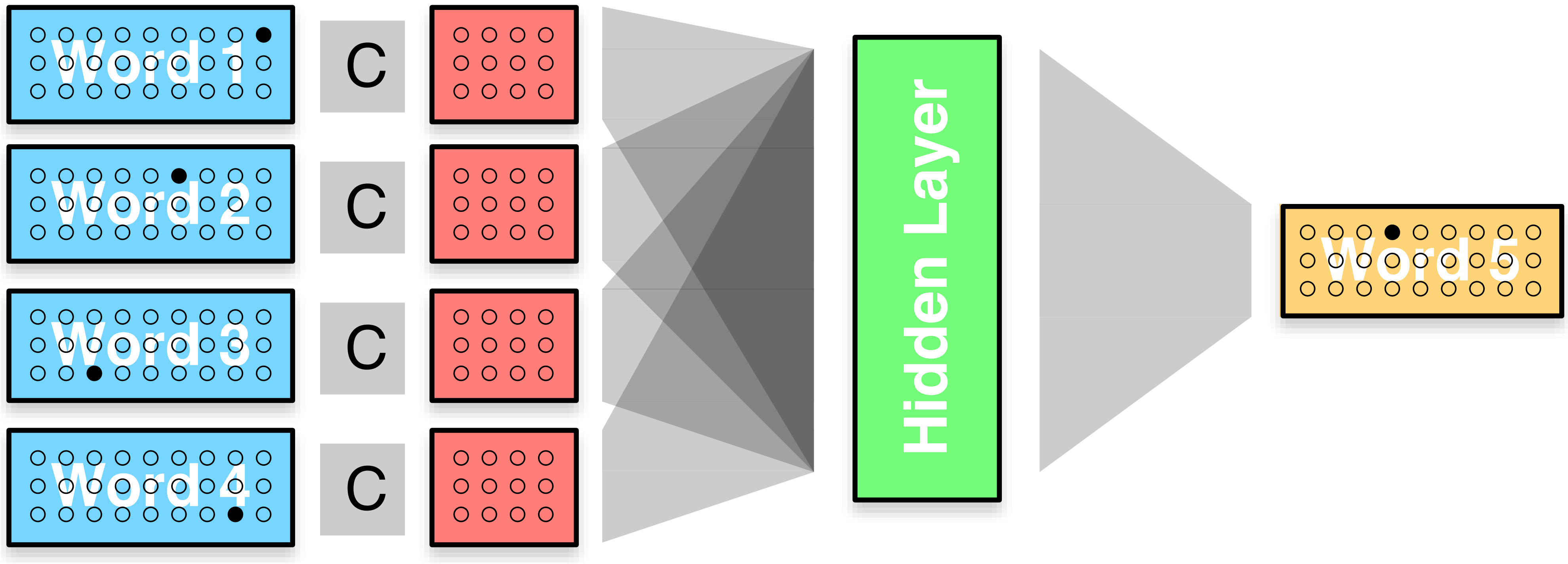
Are Word Embeddings Magic?



- Morphosyntactic regularities (Mikolov et al., 2013)
  - adjectives base form vs. comparative, e.g., good, better
  - nouns singular vs. plural, e.g., year, years
  - verbs present tense vs. past tense, e.g., see, saw

- Semantic regularities
  - clothing is to shirt as dish is to bowl
  - evaluated on human judgment data of semantic similarities

Add a Hidden Layer

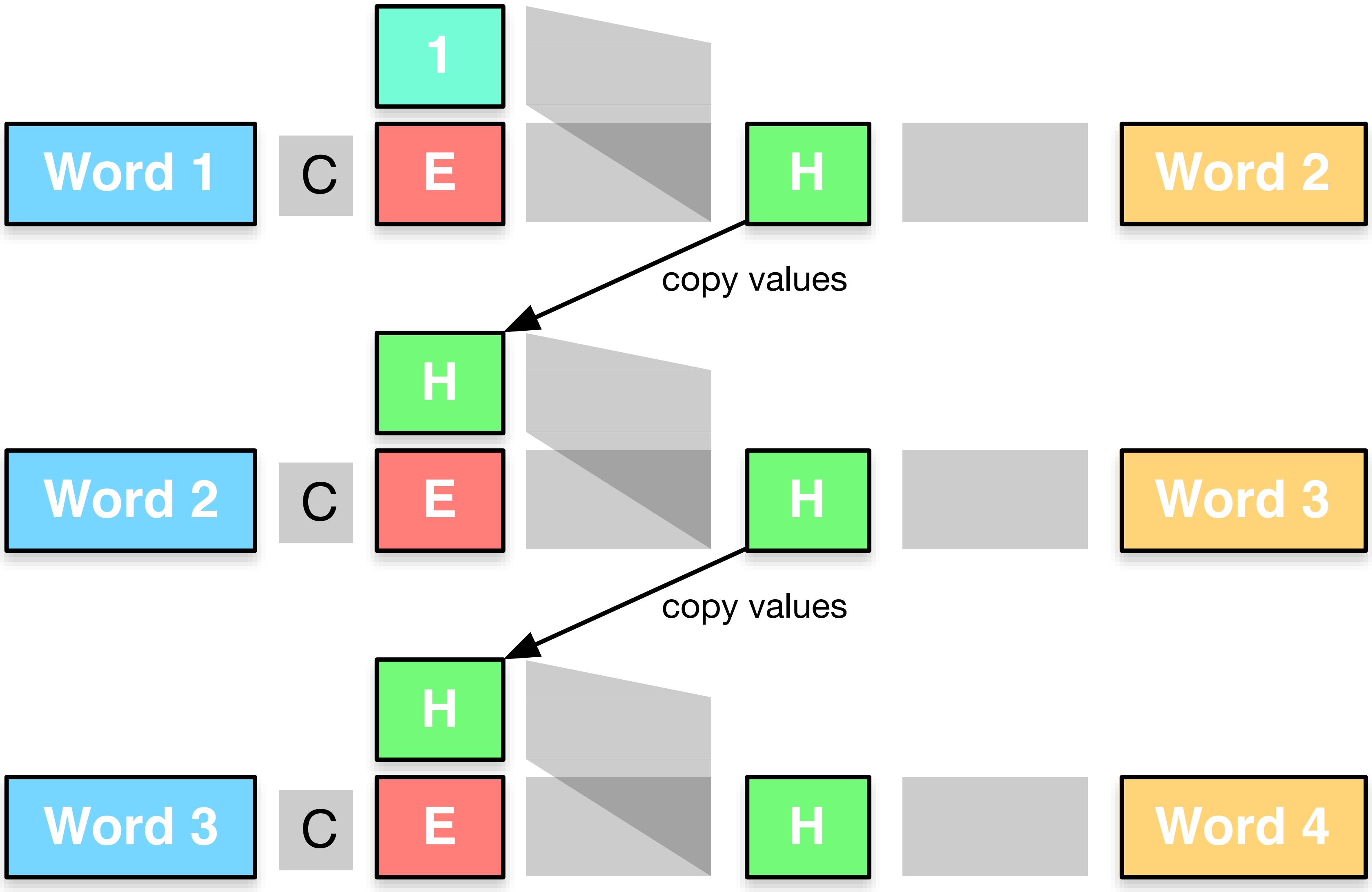


- Map each word first into a lower-dimensional real-valued space
- Shared weight matrix C

Details (Bengio et al., 2003)

- Add direct connections from embedding layer to output layer
- Activation functions
  - input→embedding: none
  - embedding→hidden: tanh
  - hidden→output: softmax
- Training
  - loop through the entire corpus
  - update between predicted probabilities and 1-hot vector for output word

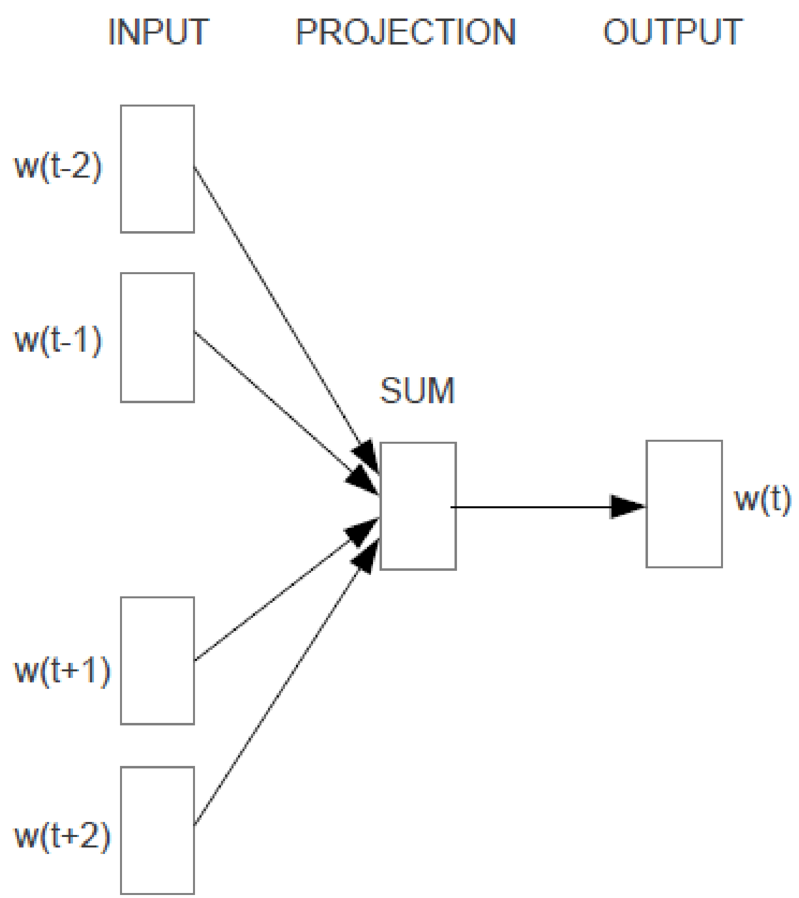
Recurrent Neural Networks



Word2vec

- Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, 2013
- Computationally efficient creation of word embeddings
- Evaluated on syntactic and semantic word similarity
- Pre-trained word embeddings created with these methods can be used in many contexts
  - Including neural machine translation

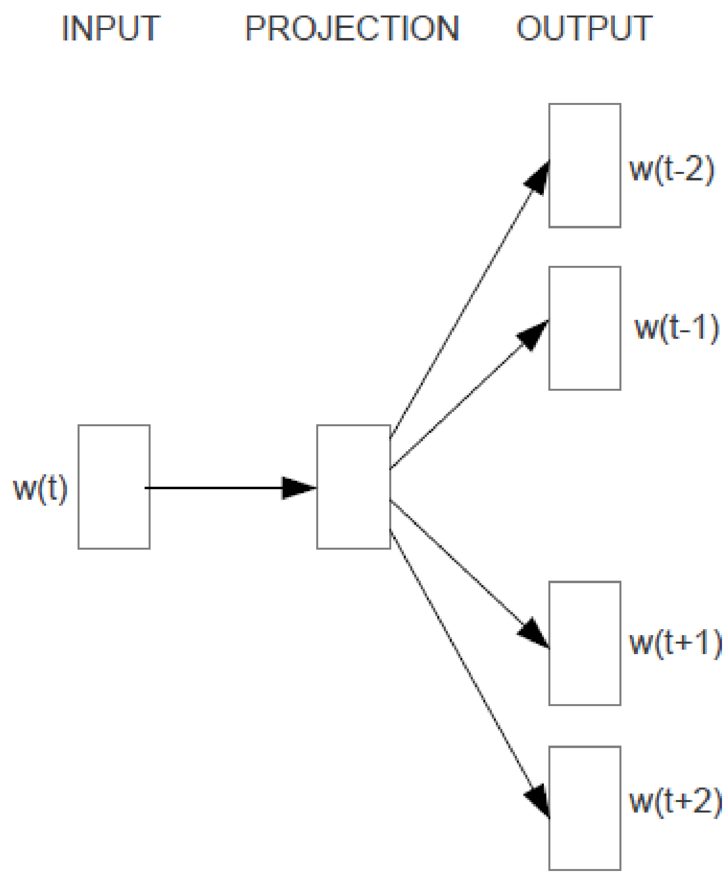
Continuous bag-of-words



CBOW

Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, 2013

Skip-gram



Skip-gram

Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, 2013

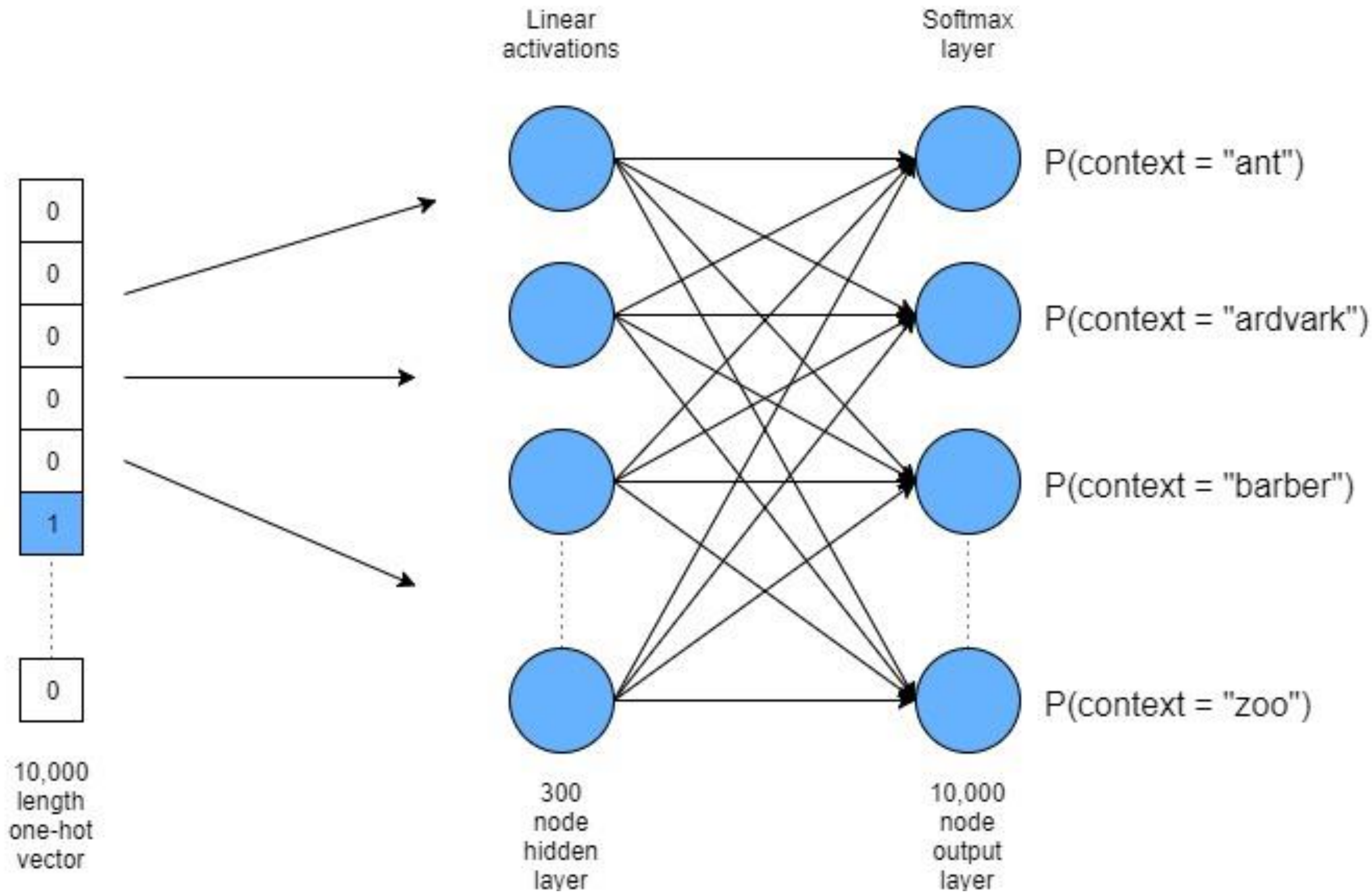
Examples of Learned Relationships

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

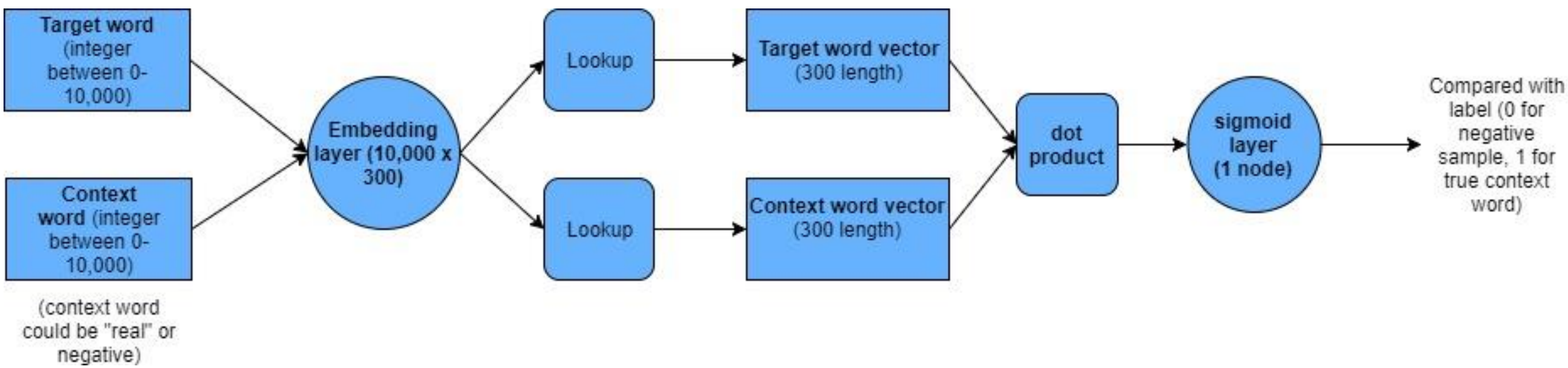
Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, 2013

Training Word2Vec Models



Credit:http://adventuresinmachinelearning.com/word2vec-keras-tutorial/

Training skip-gram models with negative sampling



Credit:http://adventuresinmachinelearning.com/word2vec-keras-tutorial/

HW5 Training Data

- Original training data fra.txt
  - 149861 sentences
  - Tab-separated
  - Sorted by increasing sentence length
    - Keep this in mind when looking at training and validation loss
- Split randomly, but in order into
  - 148861 sentence pairs fr_en.train.txt
  - 1000 sentence pairs fr_en.test.txt
    - First 500 sentences: fr_en.test_small.txt
- Data should probably be shuffled to be more realistic
  - Allows for incrementally adding longer data though

HW5: Non-coding Improvement Suggestions

- Training the system for more (or less) epochs (command line parameter --epochs)
- Training the system with more training data (command line parameter --num-samples).
- Training with a different Embedding dimension (no command line parameter, variable embedding_dim)
- Training with a different LSTM layer dimension (no command line parameter, variable latent_dim)
- Lowercase the training data
- Pre-process the training data with a different tokenizer
- Shuffle training data

HW5: Coding Improvement Suggestions

- Add a dropout layer to avoid over-fitting to the training data
- Add additional LSTM layers
- Reverse the input sentence
- Implement beam decoding
- Use pre-trained word embeddings

HW5: Using Pre-trained Word Embeddings

- How to do this in Keras
  - http://freecontent.manning.com/deep-learning-for-text/
- French word embeddings
  - E.g. http://fauconnier.github.io

References

- Axelrod et. al., Domain Adaptation via Pseudo In-Domain Data Selection, 2011, EMNLP
- Servan et. al., Domain specialization: a post-training domain adaptation for Neural Machine Translation, 2016
- Mikolov et. al., Efficient Estimation of Word Representations in Vector Space, 2013
- Mikolov et. al., Distributed Representations of Words and Phrases and their Compositionality, 2013