Adaptation

Philipp Koehn

27 October 2020

Adaptation

- Better quality when system is adapted to a task
- Domain adaptation to a specific domain, e.g., information technology
- Some training more relevant
- May also adapt to specific user (personalization)
- May optimize for a specific document or sentence

domains

Domain

- Definition

  a collection of text with similar topic, style, level of formality, etc.

- Practically: a corpus that comes from a specific source

Example

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**

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

Differences in Corpora

Medical 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

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.

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!

Dimensions

Topic The subject matter of the text, such as politics or sports.

Modality How was this text originally created? Is this written text or transcribed speech, and if speech, is it a formal presentation or an informal dialogue full of incompleted and ungrammatical sentences?

Register Level of politeness. In some languages, this is very explicit, such as the use of the informal *Du* or the formal *Sie* for the personal pronoun *you* in German.

Intent Is the text a statement of fact, an attempt to persuade, or communication between multiple parties?

Style Is it a terse informal text, are full of emotional and flowery language?

Dimensions

- In reality, no clear information about dimensions
- For example: Wikipedia
  - spans a whole range of topics
  - fairly consistent in modality and style

Dimensions

- In reality, no clear information about dimensions

- For example: Wikipedia
  - spans a whole range of topics
  - fairly consistent in modality and style

- Practical goal: enforce a certain level of politeness

- Probably
  - European parliament proceedings more polite
  - movie subtitles less polite

Impact of Domain

- Different word meanings
  - *bat* in baseball
  - *bat* in wildlife report

Impact of Domain

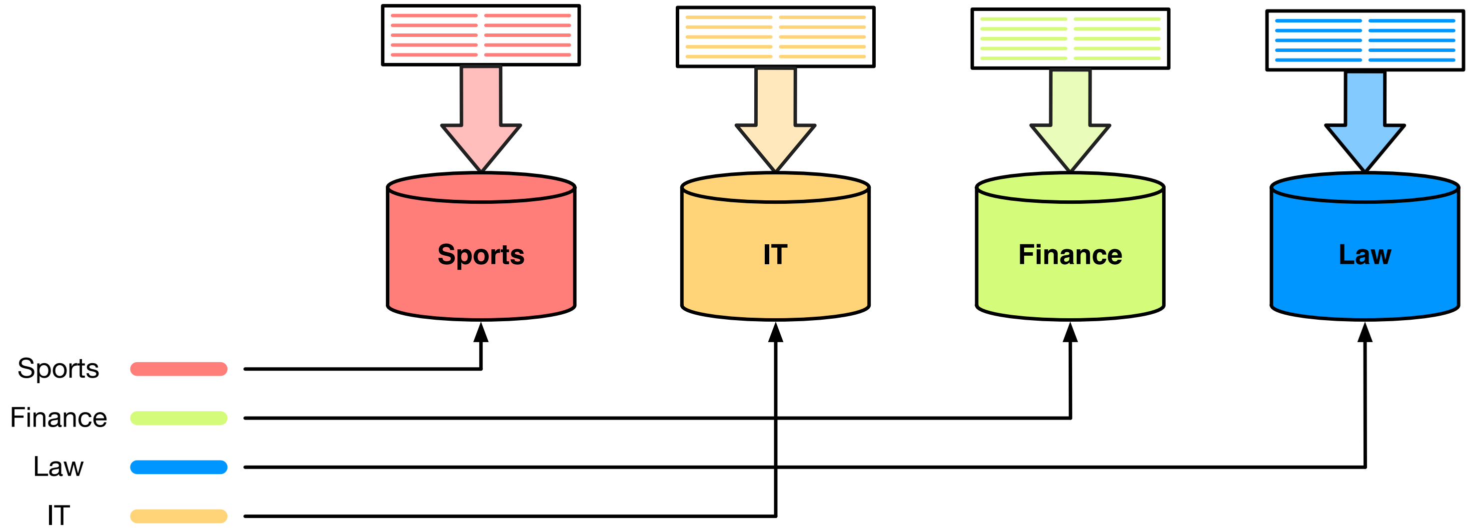
- Different word meanings
  - *bat* in baseball
  - *bat* in wildlife report
- Different style
  - *What's up, dude?*
  - *Good morning, sir.*

Diverse Problem

- Data may differ narrowly or drastically
- Amount of relevant and less relevant data differ
- Data may be split by domain or mixed
- Data may differ by quality
- Each corpus may be relatively homogeneous or heterogeneous
- May need to adapt on the fly

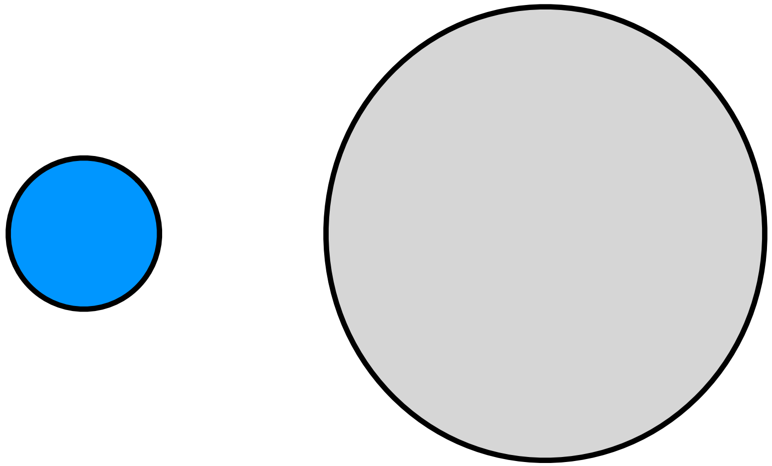
$\Rightarrow$ Different methods may apply, experimentation needed

Multiple Domain Scenario



- Multiple collections of data, clearly identified
  e.g., sports, information technology, finance, law, ...
- Train specialized model for each domain
- Route test sentences to appropriate model (using classifier, if not known)
- Probabilistic assignment

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

Adaptation Effects

German source *Verfahren und Anlage zur Durchführung einer exothermen Gasphasenreaktion an einem heterogenen partikelförmigen Katalysator*

Human reference translation *Method and system for carrying out an exothermic gas phase reaction on a heterogeneous particulate catalyst*

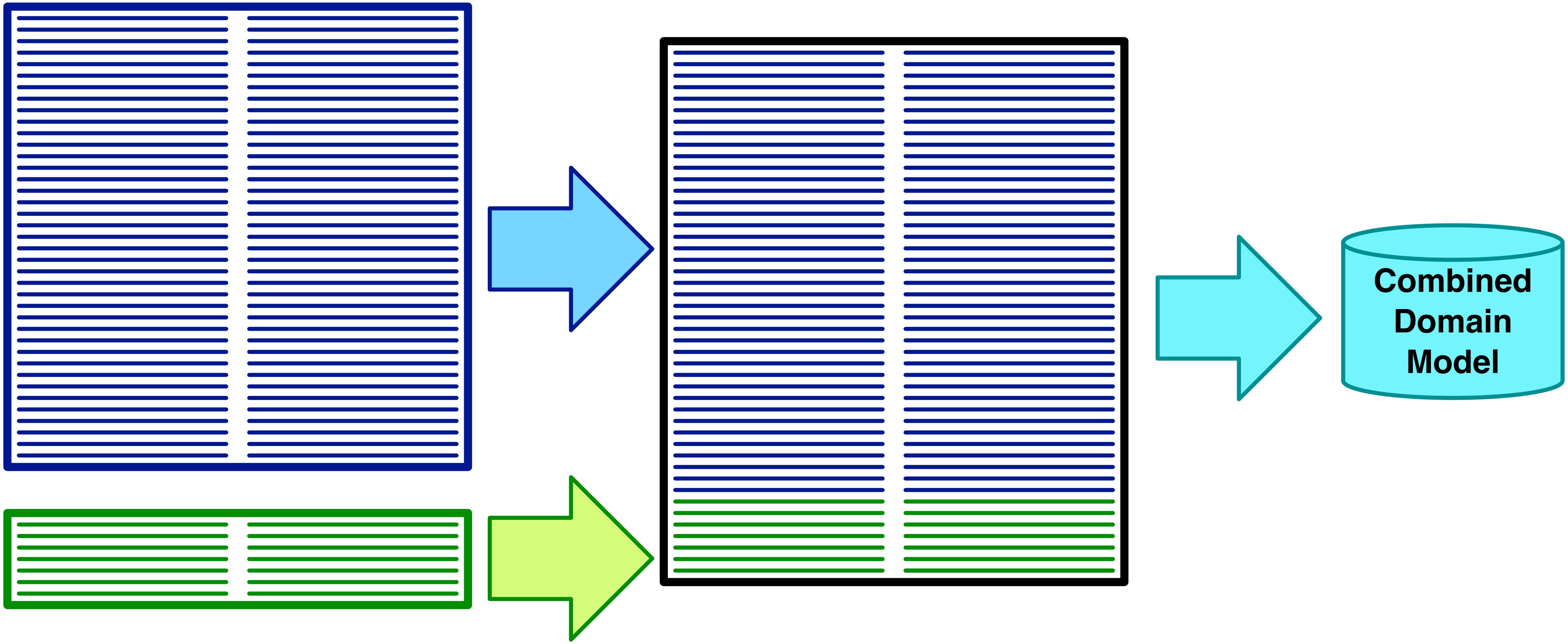
General model translation *Procedures and equipment for the implementation of an exothermen gas response response to a heterogeneous particle catalytic converter*

In-Domain (chemistry patents) model translation *Method and system for carrying out an exothermic gas phase reaction on a heterogeneous particulate catalyst*

- Stylistic, e.g., *method*, *system* vs. *procedures*, *equipment*)
- Word sense, e.g., *catalyst* vs. *catalytic converter*)
- Better language coverage
  e.g., *exothermic gas phase reaction* vs. *exothermen gas response response*

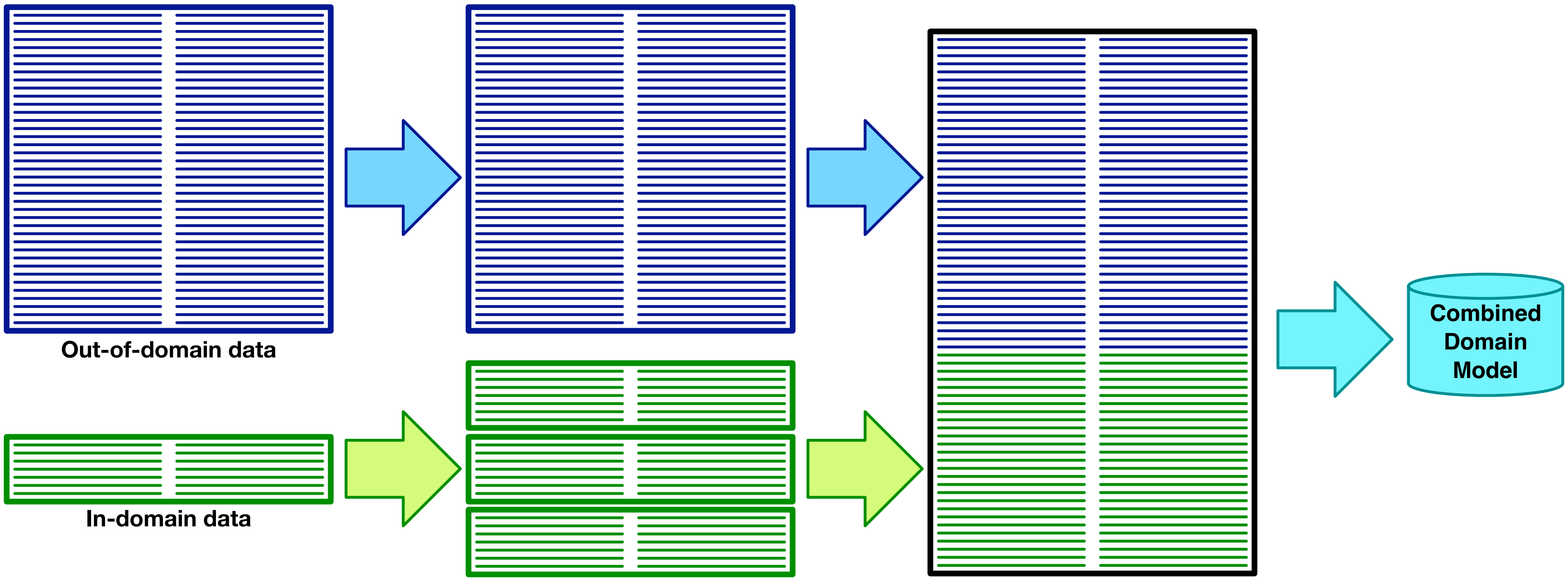
mixture models

Combine Data



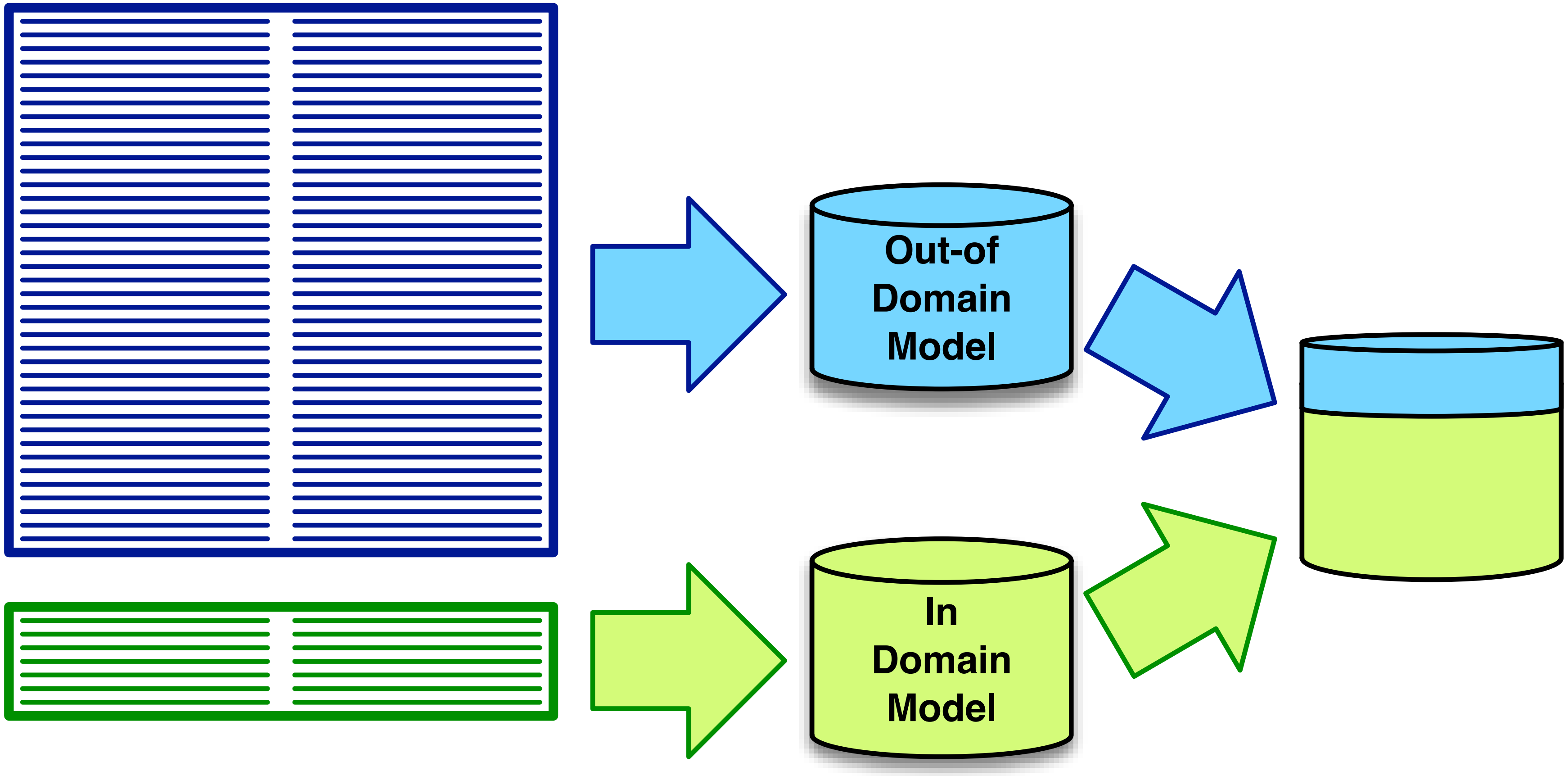
- Too biased towards out of domain data
- May flag translation options with indicator feature functions

Interpolate Data



Oversample in-domain data

Interpolate Models



Domain-Aware Training

- Train a model on all domains
- Indicate domain for each input sentence
- Domain token
  - append domain token to each input sentence, e.g., <SPORTS>
  - label training data
  - label test data
- Neural machine translation models
  - domain token will have word embedding
  - attention model will rely on domain token as needed

Unknown Domain at Test Time

- Domain of input sentence unknown
- Classifier: predict domain of input sentence
  - predict domain token
  - augment input sentence
- Probability distribution over domains

Unknown Domain at Test Time

- Domain of input sentence unknown

- Classifier: predict domain of input sentence
  - predict domain token
  - augment input sentence

- Probability distribution over domains
  - sentences may not fall neatly into one of our pre-defined domains
  - e.g., rule violation in sports → SPORTS, LAW
  - encode soft domain assignment in vector
  - may be also used to label training data

Fine-Grained Domains: Personalization

- Thousands of domains
  - machine translation system personalized for individual translators
  - machine translation system optimized for authors/speakers
- Domain token/classification idea does not scale well
- Not much data for each domain

Fine-Grained Domains: Personalization

- Only influence word prediction layer

- Recall output word distribution t_i as a softmax given
  - previous hidden state (s_{i-1})
  - previous output word embedding (Ey_{i-1})
  - input context (c_i)

$$t_i = \text{softmax}\big(W(U s_{i-1} + V E y_{i-1} + C c_i) + b\big)$$

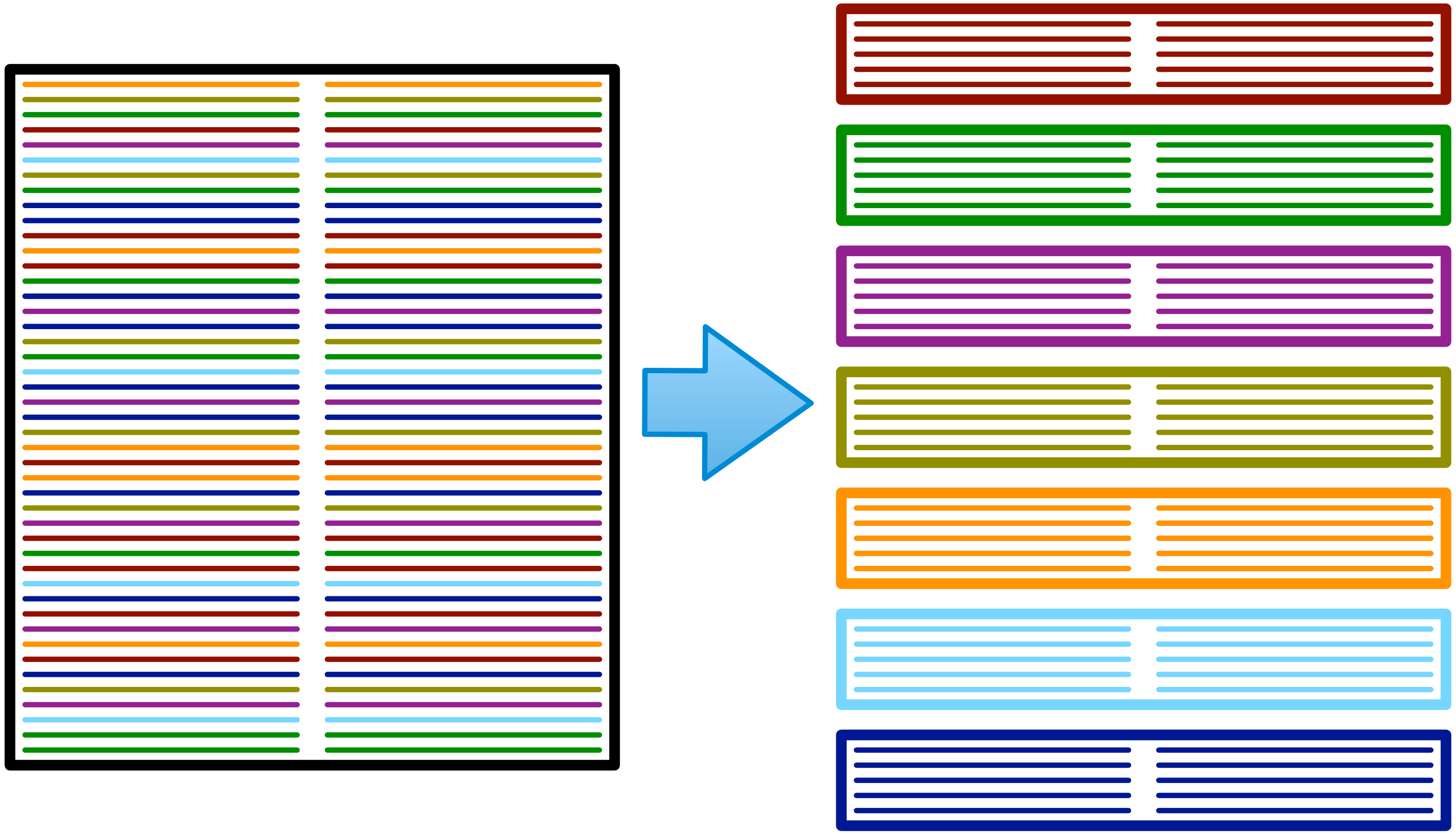
- More generally, prediction given some conditioning vector z_i

$$t_i = \text{softmax}\big(W z_i + b\big)$$

- Add an additional bias term β_p specific to a person p

$$t_i = \text{softmax}\big(W z_i + b + \beta_p\big)$$

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)

Latent Dirichlet Allocation (LDA)

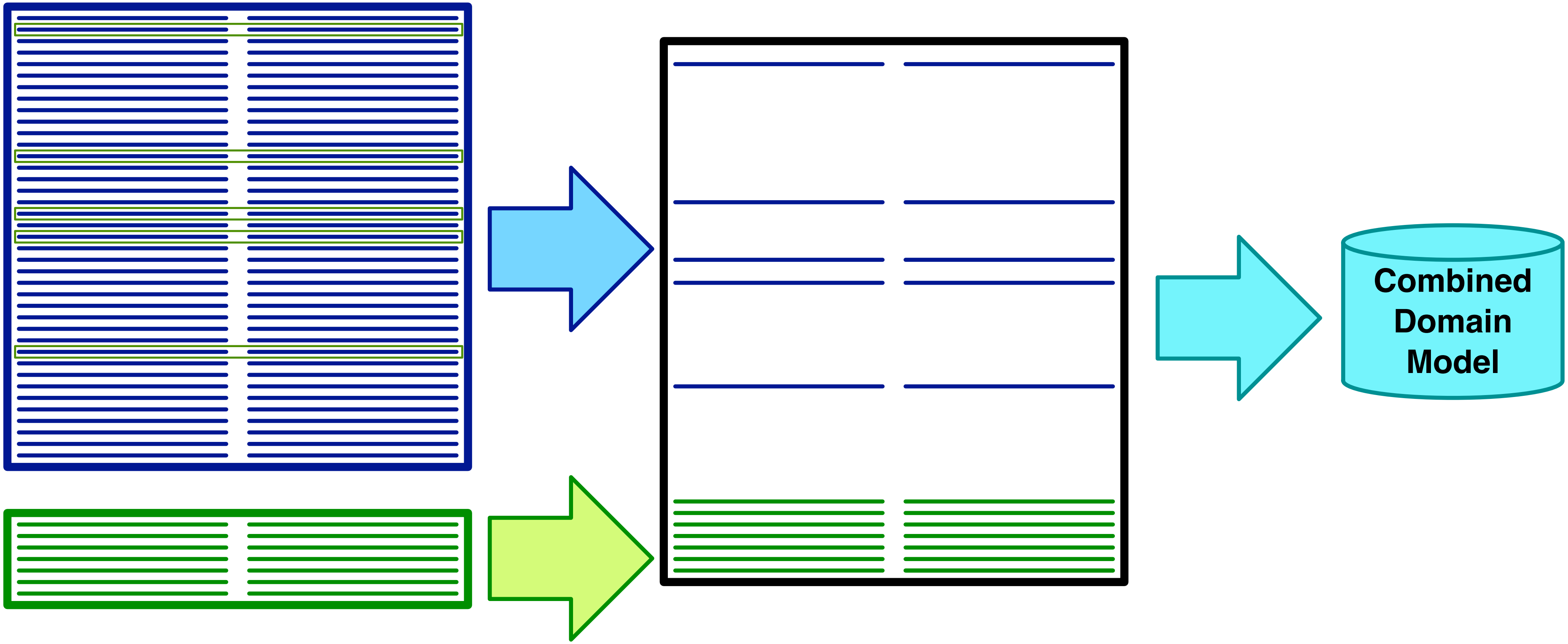
- Formalized as a graphical model
- Sentences belong to a fixed number of topics
- Model
  - predicts distribution over topics
  - predicts words based on each topic
- For instance, typical topics
  - *European, political, policy, interests, ...*
  - *crisis, rate, financial, monetary, ...*

Sentence Embeddings

- Sentence embeddings
  - simple method: average of embedding of the words in the sentence
  - ongoing research on more complex methods
- Cluster sentences into topics: k-means clustering
  - randomly generate centroids (vectors in sentence embedding space)
  - assign each sentence to its closest centroid
  - re-compute centroid as center of the embeddings of its assigned sentences
  - iterate
- Input sentence to be translated
  - assign to topic, based on proximity to centroids
  - translate with topic-specific model

subsampling

Sentence Selection

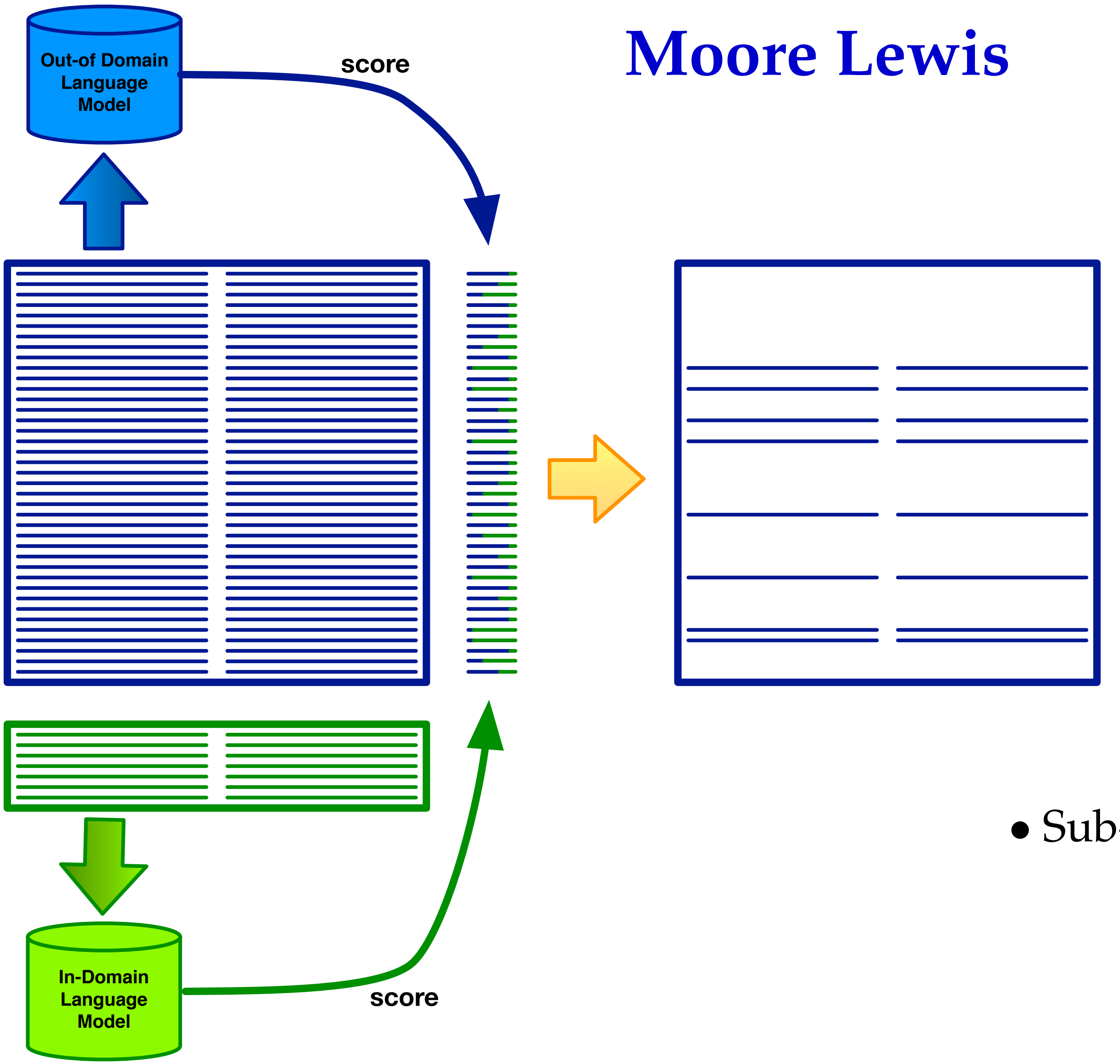


- Select out-of-domain sentence pairs that are similar to in-domain data

Sentence Selection

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

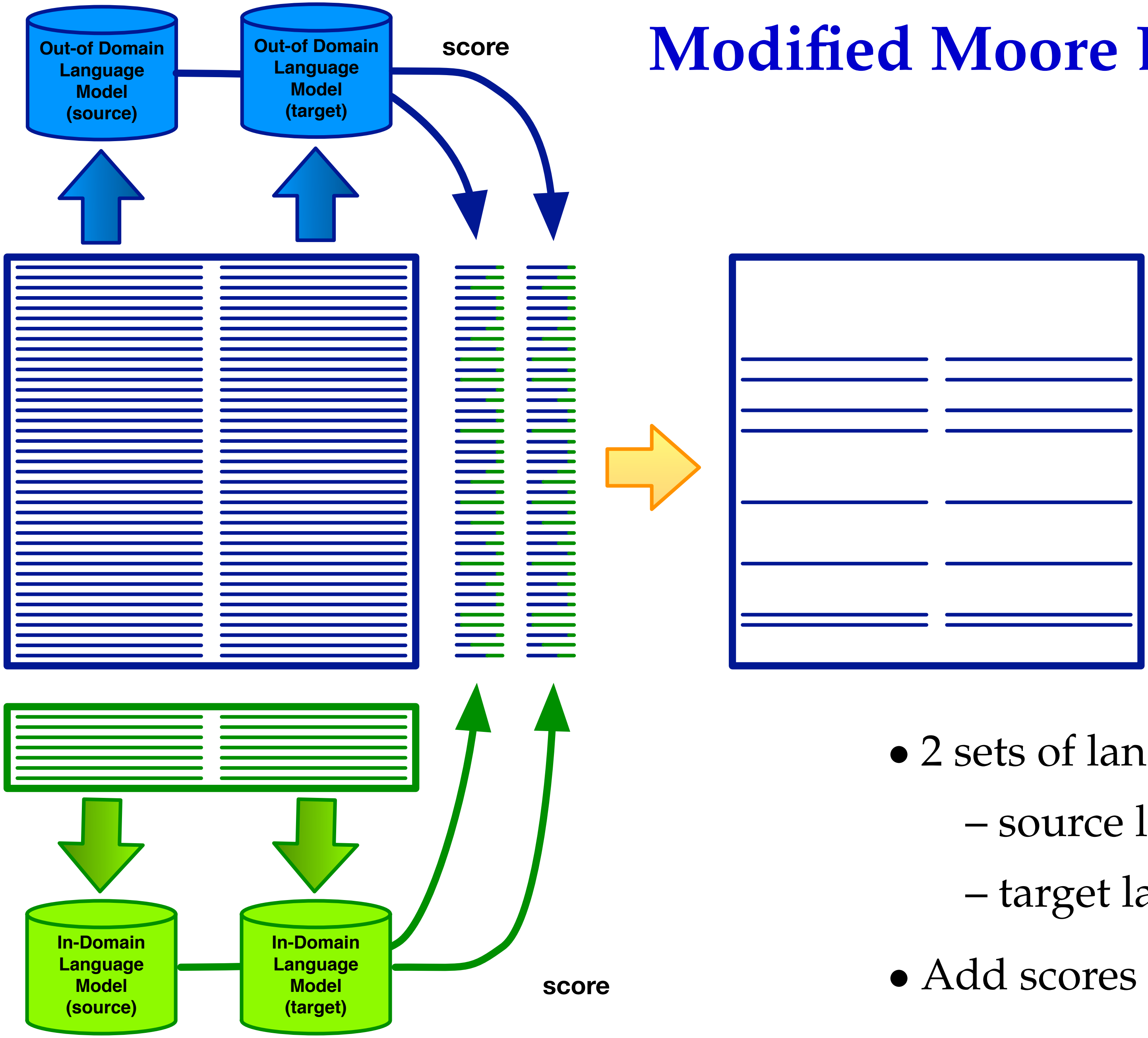
Moore Lewis



- 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?

Coverage-Based Methods

- Problem with subsampling sentences based on similarity: not much new is added

- Original goal: increase coverage with out-of-domain data

$\rightarrow$ coverage-based selection

Basic Approach

- Score each candidate sentence pair to be added based on word-based score

$$\frac{1}{|s_i|}\sum_{w \in s} \text{score}(w, s_{1,..,i-1})$$

- Simple word score: check if word w occurred in the previously added sentences $s_1, ..., s_{i-1}$

$$\text{score}(w, s_{1,..,i-1}) = \begin{cases} 0 & \text{if } w \in s_1, ..., s_{i-1} \\ 1 & \text{otherwise} \end{cases}$$

- Add sentence with highest score

Scoring N-Grams

- Compute coverage of n-grams, not just words

$$\frac{1}{|s_i| \times N} \sum_{n=0}^{N-1} \sum_{w_{j,\ldots,j+n} \in s} \text{score}(w_{j,\ldots,j+n}, s_{1,..,i-1})$$

Feature Decay

- Not hard 0/1 scoring
- Decaying function based on frequency

$$\text{score}(w, s_{1,..,i-1}) = \text{frequency}(w, s_{1,..,i-1})\ e^{-\lambda\ \text{frequency}(w, s_{1,..,i-1})}$$

- May also consider frequency of n-grams in raw corpus
  (avoid overfitting to rare n-grams)

Instance Weighting

- So far: either include sentence pair or not
- Now: weigh sentence pair based on relevance

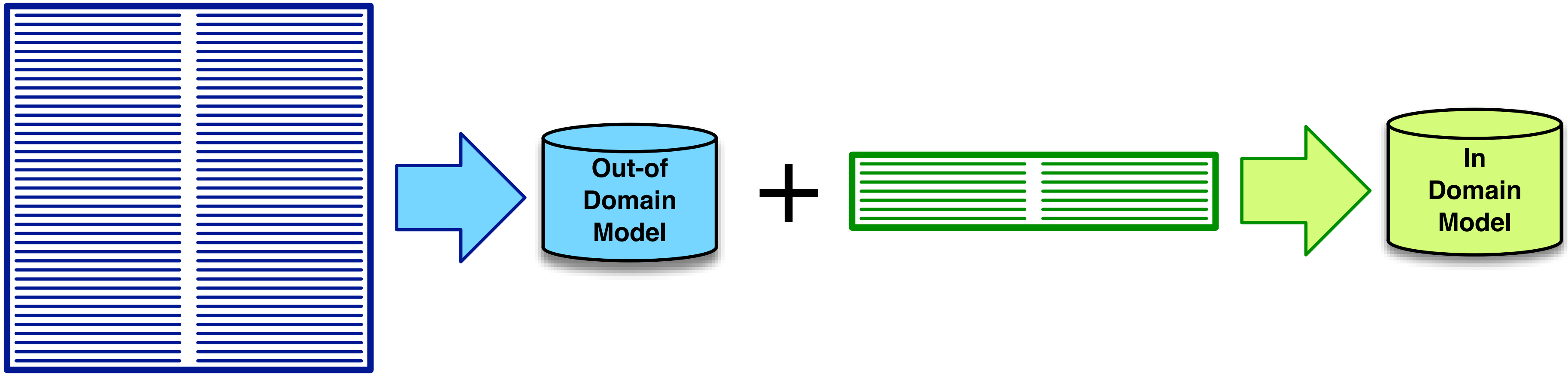
Instance Weighting

- So far: either include sentence pair or not
- Now: weigh sentence pair based on relevance

- Use same scoring metrics as previously for filtering
- Scale learning rate by relevance score

fine tuning

Fine-Tuning



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

Catastrophic Forgetting

- Fine tuning may overfit to in-domain data (catastrophic forgetting)
- Two goals
  - do well on in-domain data
  - maintain quality on out-of-domain data
- Makes model more robust on in-domain data as well

Updating only Some Model Parameters

- Too many parameters, too few in-domain data
- Update only some parameters
  - weights for decoder state progression
  - output word prediction softmax
  - output word embeddings

Adaptation Parameters

- Leave general model parameters fixed
- Learning hidden unit contribution (LHUC) layer
  - learn scaling values in narrow range (say, factor 0 to 2)

$$a(\rho) = \frac{2}{1 + e^{\rho}}$$

  - scale values of decoder state s.

$$s_{\text{LHUC}} = a(\rho) \circ s$$

- Can be easily turned off

Regularized Training Objective

- Stated goal: do not diverge too far from the original model

- Default training objective
  - reduce the error on word predictions probability $t_i[y_i]$
  - given to the correct output word y_i at time step i

$$\text{cost} = -\log t_i[y_i]$$

Regularized Training Objective

- Stated goal: do not diverge too far from the original model
- Default training objective
  - reduce the error on word predictions probability $t_i[y_i]$
  - given to the correct output word y_i at time step i

$$\text{cost} = -\log t_i[y_i]$$

- Measurement of difference to general model's prediction t_i^{BASE}

$$\text{cost}_{\text{REG}} = \sum_{y \in V} t_i^{\text{BASE}}[y] \log t_i[y]$$

Regularized Training Objective

- Stated goal: do not diverge too far from the original model

- Default training objective
  - reduce the error on word predictions probability $t_i[y_i]$
  - given to the correct output word y_i at time step i

$$\text{cost} = -\log t_i[y_i]$$

- Measurement of difference to general model's prediction t_i^{BASE}

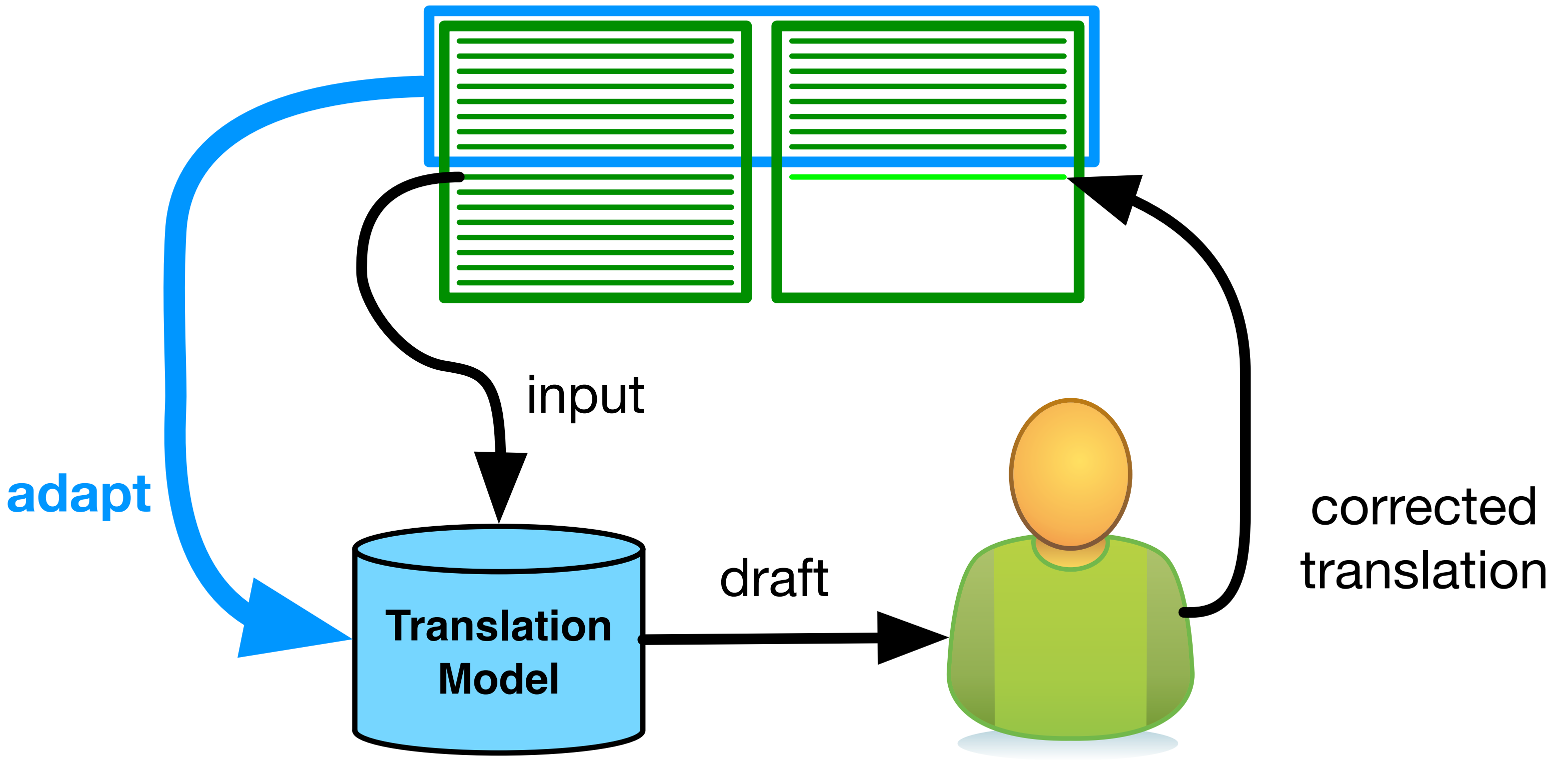
$$\text{cost}_{\mathsf{REG}} = \sum_{y \in V} t_i^{\mathsf{BASE}}[y] \log t_i[y]$$

- Combine both training objectives

$$(1-\alpha)\ \text{cost} + \alpha\ \text{cost}_{\mathsf{REG}}$$

- Balancing factor α can be used to balance in-domain / out-of-domain quality

Document-Level Adaptation



- Computer aided translation: translator post-edits machine translation
- Provides additional training data (translated sentences)
- Incrementally update model

Sentence-Level Adaptation

- Adapt model to each sentence to be translated

Sentence-Level Adaptation

- Adapt model to each sentence to be translated

- Find most similar sentence in parallel corpus (fuzzy match)

- Retrieve it and its translation

- Adapt model with this sentence pair

Curriculum Training

- Recall: relevance score for each sentence pair
- Training epochs
  - start with all data (100%)
  - train only on somewhat relevant data (50%)
  - train only on relevant data (25%)
  - train only on very relevant data (10%)