Multilingual / Cross-lingual Methods

Wei Xu

(many slides from Greg Durrett)

Announcements

- This is the last class.
- Final Project presentations on Apr 29 2:40pm (final exam time)
- Course Instructor Opinion Surveys (CIOS): please fill these out

Frontiers in MT

Low-Resource MT

- Particular interest in deploying MT systems for languages with little or no parallel data

- BPE allows us to transfer models even without training on a specific language

- Pre-trained models can help further

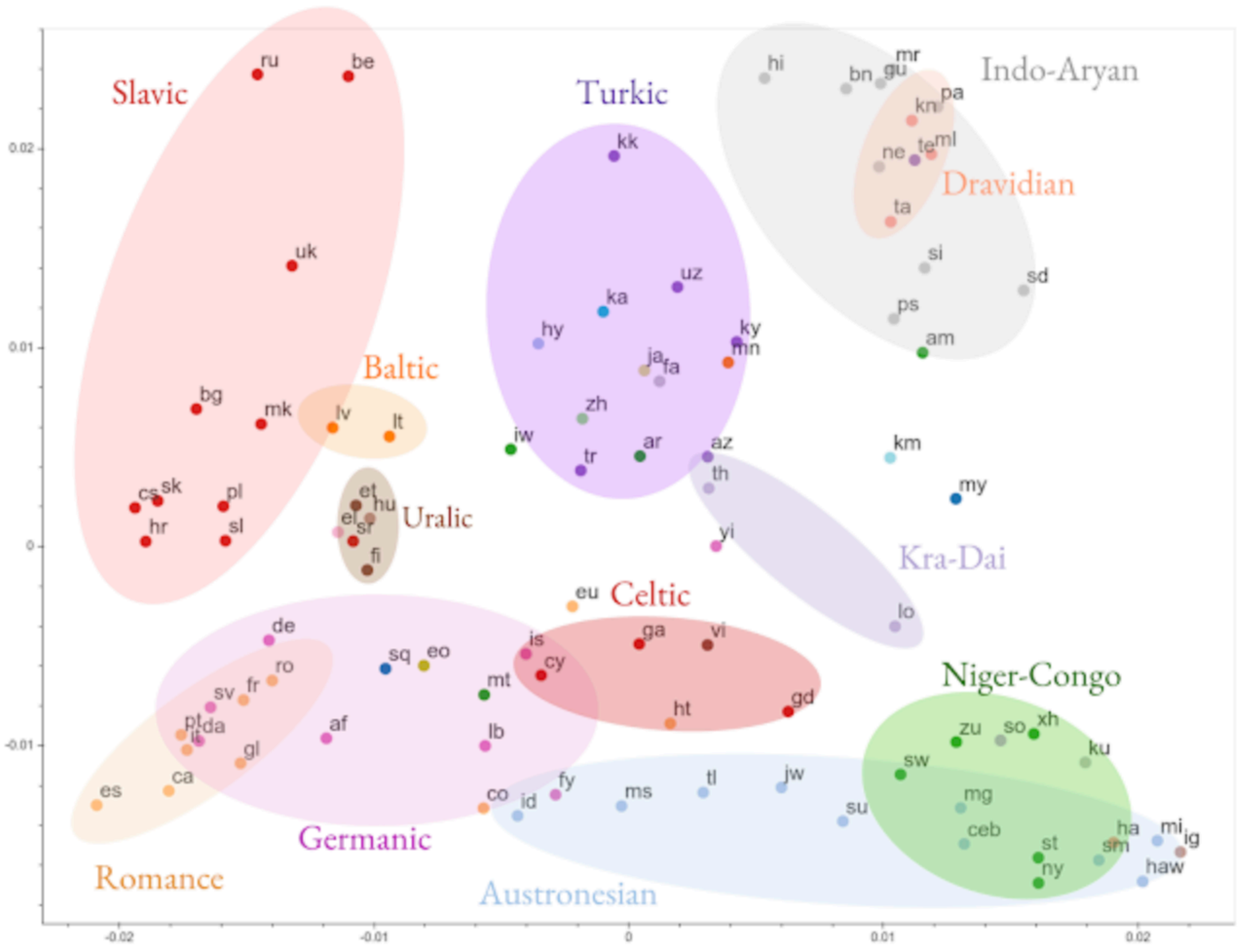
Burmese, Indonesian, Turkish

Transfer	BLEU My→En	BLEU Id→En	BLEU Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

Table 3: Investigating the model's capability to restore its quality if we reset the parameters. We use En→De as the parent.

Aji et al. (2020)

Massively Multilingual MT



Visualization of the clustering of the encoded representations of all 103 languages, based on representational similarity. Languages are color-coded by their linguistic family.

- For 103 languages

https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

Arivazhagan et al. (2019), Kudugunta et al. (2019)

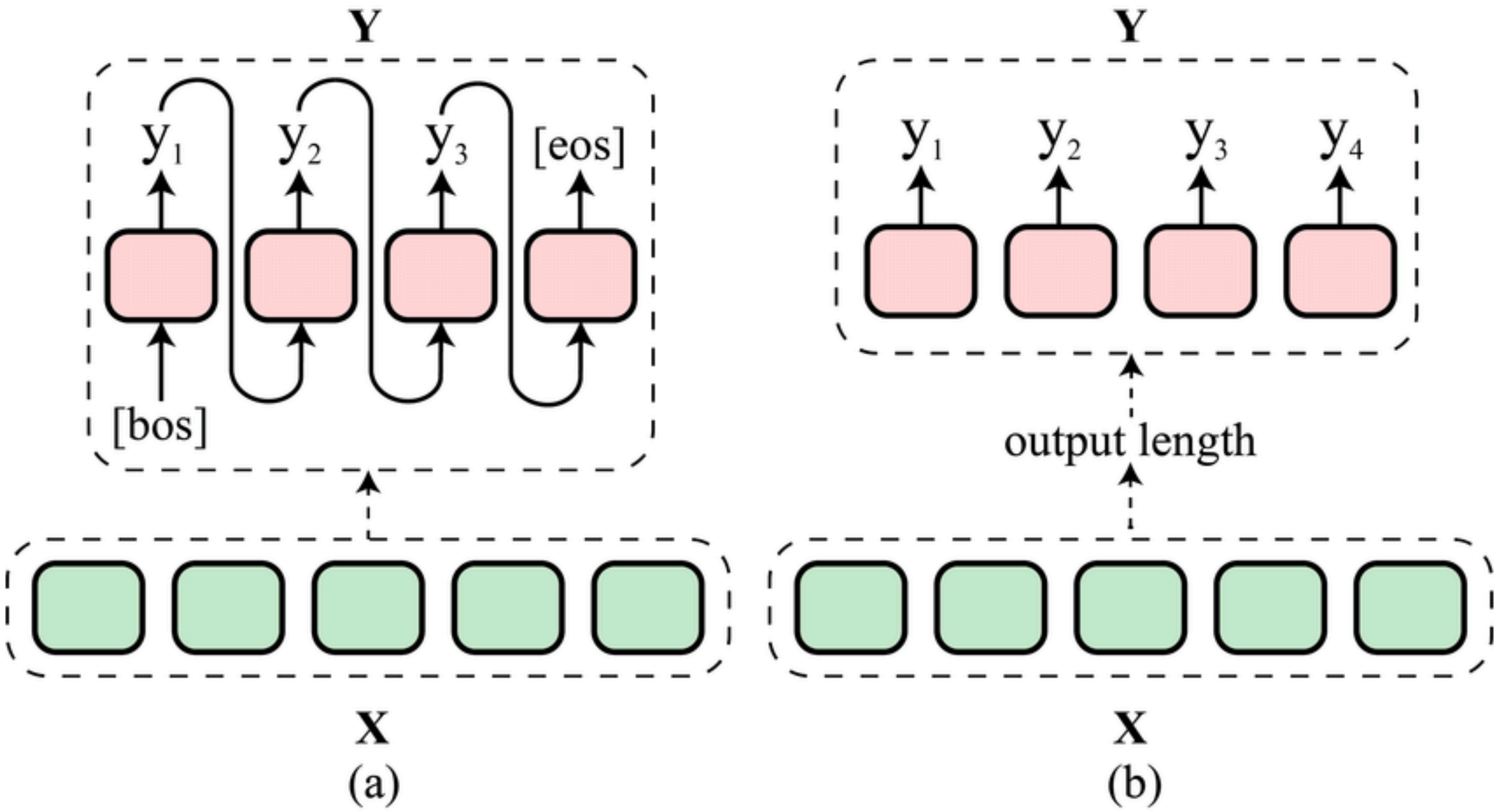
Unsupervised MT

Approach	Train/Val	Test	Loss
Supervised MT	L1-L2	L1-L2	$\mathcal{L}^{MT}_{x\rightarrow y} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X},\mathcal{Y})}\left[-\log p_{x\rightarrow y}(\mathbf{y}\|\mathbf{x})\right]$
Unsupervised MT	L1, L2	L1-L2	$\mathcal{L}^{BT}_{x\leftrightarrow y} = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\left[-\log p_{y\rightarrow x}(\mathbf{x}\|g^*(\mathbf{x}))\right] + \mathbb{E}_{\mathbf{y}\sim\mathcal{Y}}\left[-\log p_{x\rightarrow y}(\mathbf{y}\|h^*(\mathbf{y}))\right]$ g^*, h^*: sentence predictors

- Common principles of unsupervised MT
  - Language models
  - (Iterative) Back-translation!

Lample et al. (2018)

Non-Autoregressive NMT



- Q: why non-autoregressive? Pros and cons?

https://homes.cs.washington.edu/~jkasai/2020-01-28/nat/

Gu et al. (2018), Ghazvininejad et al. (2019), Kasai et al. (2020)

Efficiency of NMT

**SIXTH CONFERENCE ON
MACHINE TRANSLATION (WMT21)**

**November 10-11, 2021
Punta Cana (Dominican Republic) and Online**

Shared Task: Efficiency

[HOME] [SCHEDULE] [PAPERS] [AUTHORS] [RESULTS]
TRANSLATION TASKS: [NEWS] [SIMILAR LANGUAGES] [BIOMEDICAL] [EUROPEAN LOW RES MULTILINGUAL] [LARGE-SCALE MULTILINGUAL] [TRIANGULAR MT]
[EFFICIENCY] [TERMINOLOGY] [UNSUP AND VERY LOW RES] [LIFELONG LEARNING]
EVALUATION TASKS: [QUALITY ESTIMATION] [METRICS]
OTHER TASKS: [AUTOMATIC POST-EDITING]

Efficiency Task

The efficiency task measures latency, throughput, memory consumption, and size of machine translation on CPUs and GPUs. Participants provide their own code and models using standardized data and hardware. This is a continuation of the WNGT 2020 Efficiency Shared Task.

Multilinguality

NLP in other languages

- Other languages present some challenges not seen in English at all!
- Some of our algorithms have been specified to English
  - Neural methods are typically tuned to English-scale resources, may not be the best for other languages where less data is available
- Question:

1) What other phenomena / challenges do we need to solve?

2) How can we leverage existing resources to do better in other languages without just annotating massive data?

This Lecture

- Morphological richness: effects and challenges
- Morphology tasks: analysis, inflection, word segmentation
- Cross-lingual tagging and parsing
- Cross-lingual word representations

Morphology

What is morphology?

- Study of how words form
- Derivational morphology: create a new *lexeme* from a base

  estrange (v) => estrangement (n)

  become (v) => unbecoming (adj)

  - May not be totally regular: enflame => inflammable
- Inflectional morphology: word is inflected based on its context

  I become / she become**s**

  - Mostly applies to verbs and nouns

Morphological Inflection

- In English: I arrive   you arrive   he/she/it arrives   [X] arrived
  we arrive   you arrive   they arrive

- In French:

		singular			plural		
		first	second	third	first	second	third
indicative		je (j’)	tu	il, elle	nous	vous	ils, elles
(simple tenses)	present	arrive /a.ʁiv/	arrives /a.ʁiv/	arrive /a.ʁiv/	arrivons /a.ʁi.vɔ̃/	arrivez /a.ʁi.ve/	arrivent /a.ʁiv/
	imperfect	arrivais /a.ʁi.vɛ/	arrivais /a.ʁi.vɛ/	arrivait /a.ʁi.vɛ/	arrivions /a.ʁi.vjɔ̃/	arriviez /a.ʁi.vje/	arrivaient /a.ʁi.vɛ/
	past historic[2]	arrivai /a.ʁi.vɛ/	arrivas /a.ʁi.va/	arriva /a.ʁi.va/	arrivâmes /a.ʁi.vam/	arrivâtes /a.ʁi.vat/	arrivèrent /a.ʁi.vɛʁ/
	future	arriverai /a.ʁi.vʁɛ/	arriveras /a.ʁi.vʁa/	arrivera /a.ʁi.vʁa/	arriverons /a.ʁi.vʁɔ̃/	arriverez /a.ʁi.vʁe/	arriveront /a.ʁi.vʁɔ̃/
	conditional	arriverais /a.ʁi.vʁɛ/	arriverais /a.ʁi.vʁɛ/	arriverait /a.ʁi.vʁɛ/	arriverions /a.ʁi.və.ʁjɔ̃/	arriveriez /a.ʁi.və.ʁje/	arriveraient /a.ʁi.vʁɛ/

Morphological Inflection

- In Spanish:

		singular			plural		
		1st person	2nd person	3rd person	1st person	2nd person	3rd person
indicative		yo	tú vos	él/ella/ello usted	nosotros nosotras	vosotros vosotras	ellos/ellas ustedes
	present	llego	llegas[tú] llegás[vos]	llega	llegamos	llegáis	llegan
	imperfect	llegaba	llegabas	llegaba	llegábamos	llegabais	llegaban
	preterite	llegué	llegaste	llegó	llegamos	llegasteis	llegaron
	future	llegaré	llegarás	llegará	llegaremos	llegaréis	llegarán
	conditional	llegaría	llegarías	llegaría	llegaríamos	llegaríais	llegarían

Noun Inflection

- Not just verbs either; gender, number, case complicate things

Declension of Kind [hide ▲]

	singular			plural	
	indef.	def.	noun	def.	noun
nominative	ein	das	Kind	die	Kinder
genitive	eines	des	Kindes, Kinds	der	Kinder
dative	einem	dem	Kind, Kinde[1]	den	Kindern
accusative	ein	das	Kind	die	Kinder

- Nominative: I/he/she, accusative: me/him/her, genitive: mine/his/hers
- Dative: merged with accusative in English, shows recipient of something

  I taught the children <=> Ich unterrichte die Kinder

  I give the children a book <=> Ich gebe den Kindern ein Buch

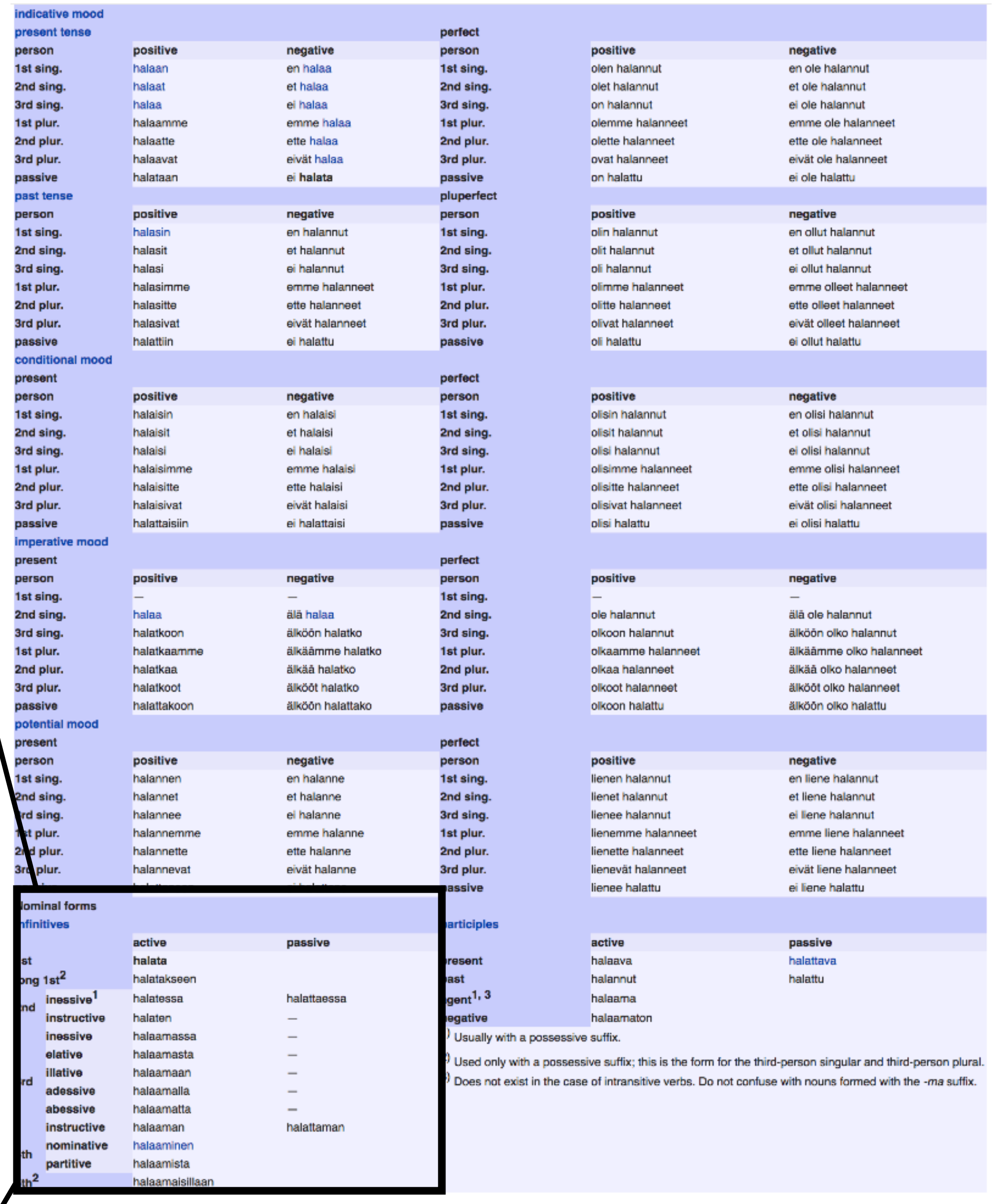
Irregular Inflection

- Common words are often irregular
  - I am / you are / she is
  - Je suis / tu es / elle est
  - Soy / está / es
- Less common words typically fall into some regular *paradigm* — these are somewhat predictable

Agglutinating Langauges

- Finnish/Hungarian (Finno-Ugric), also Turkish: what a preposition would do in English is instead part of the verb (*hug*)

		active	passive
1st		**halata**	
long 1st[2]		halatakseen	
2nd	inessive[1]	halatessa	halattaessa
	instructive	halaten	—
3rd	inessive	halaamassa	—
	elative	halaamasta	—
	illative	halaamaan	—
	adessive	halaamalla	—
	abessive	halaamatta	—
	instructive	halaaman	halattaman
4th	nominative	halaaminen	
	partitive	halaamista	
5th[2]		halaamaisillaan	



halata: “hug”

illative: “into”

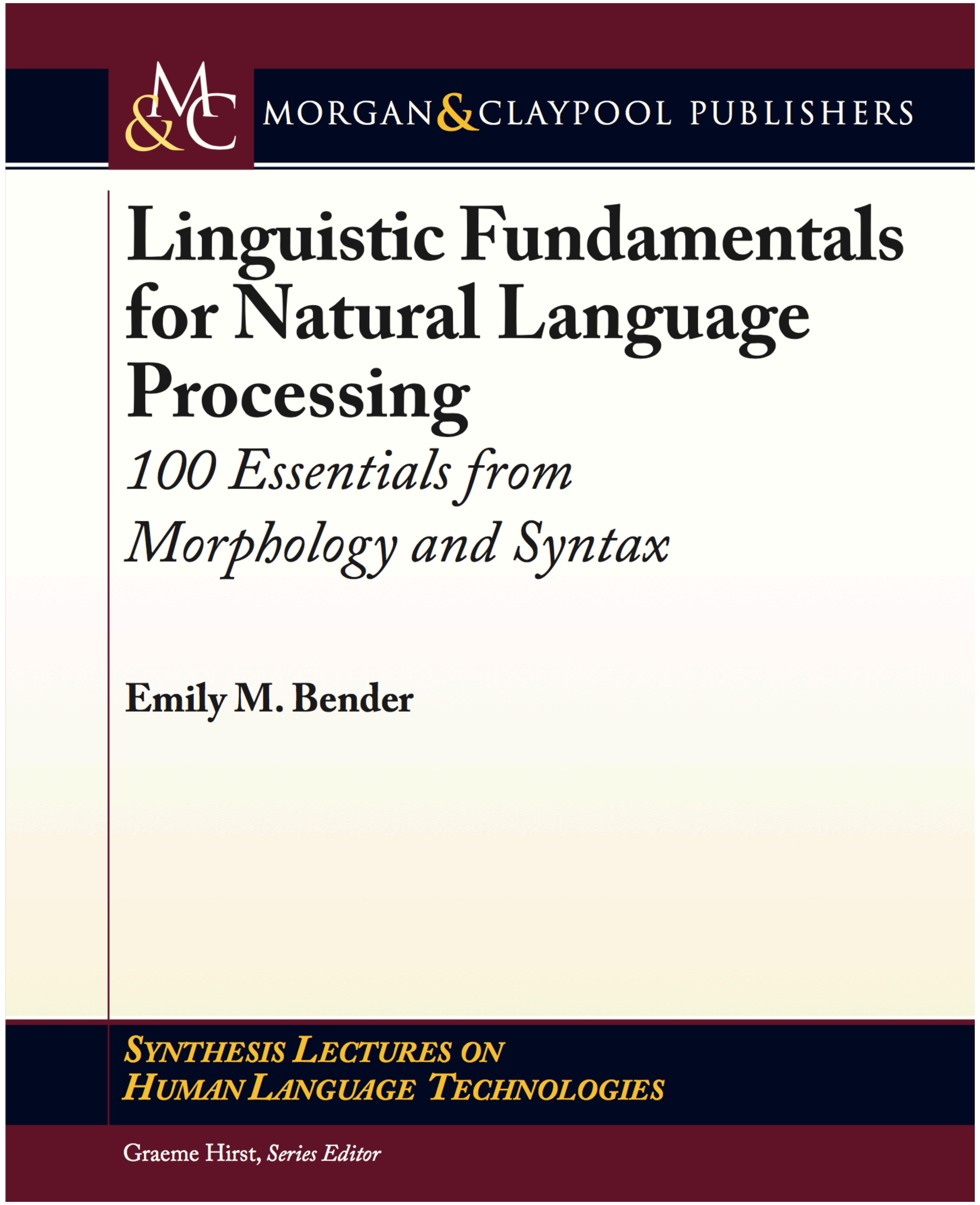
adessive: “on”

- Many possible forms — and in newswire data, only a few are observed

Morphologically-Rich Languages

- Many languages spoken all over the world have much richer morphology than English
  - CoNLL 2006 / 2007: dependency parsing + morphological analyses for ~15 mostly Indo-European languages
  - SPMRL shared tasks (2013-2014): Syntactic Parsing of Morphologically-Rich Languages
  - Universal Dependencies project (2005-now): >100 languages
- Word piece / byte-pair encoding models for MT are pretty good at handling these if there's enough data

Morphologically-Rich Languages



- Great resources for challenging your assumptions about language and for understanding multilingual models!

Morphological Analysis/Inflection

Morphological Analysis

- In English, lexical features on words and word vectors are pretty effective
- In other languages, **lots** more unseen words due to rich morphology! Affects parsing, translation, …
- When we're building systems, we probably want to know base form + morphological features explicitly
- How to do this kind of *morphological analysis*?

Morphological Analysis: Hungarian

But the government does not recommend reducing taxes.

Ám a kormány egyetlen adó csökkentését sem javasolja .

n=singular|case=nominative|proper=no

deg=positive|n=singular|case=nominative

n=singular|case=nominative|proper=no

n=singular|case=accusative|proper=no|pperson=3rd|pnumber=singular

mood=indicative|t=present|p=3rd|n=singular|def=yes

Morphological Analysis

- Given a word in context, need to predict what its morphological features are
- Basic approach: combines two modules:
  - Lexicon: tells you what possibilities are for the word
  - Analyzer: statistical model that disambiguates
- Models are largely CRF-like: score morphological features in context
- Lots of work on Arabic inflection (high amounts of ambiguity)

Morphological Inflection

- Inverse task of analysis: given base form + features, inflect the word
- Hard for unknown words — need models that generalize

w i n d e n →

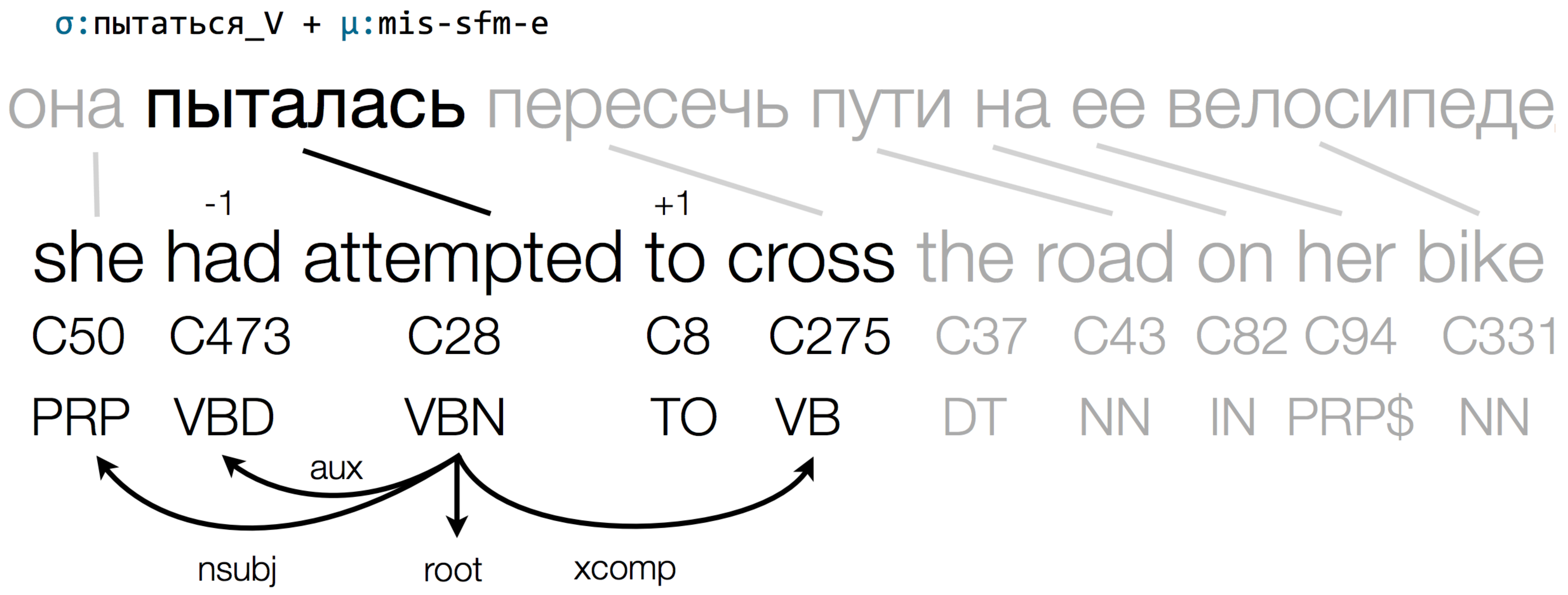
conjugation of winden [hide ▲]

infinitive	winden				
present participle	windend				
past participle	gewunden				
auxiliary	haben				
	indicative			**subjunctive**	
present	ich winde	wir **winden**	**i**	ich winde	wir **winden**
	du windest	ihr windet		du windest	ihr windet
	er windet	sie **winden**		er winde	sie **winden**
preterite	ich wand	wir wanden	**ii**	ich wände	wir wänden
	du wandest	ihr wandet		du wändest	ihr wändet
	er wand	sie wanden		er wände	sie wänden
imperative	winde (du)	windet (ihr)			

composed forms of winden [show ▼]

Durrett and DeNero (2013)

Morphological Inflection



- Machine translation where phrase table is defined in terms of lemmas
- "Translate-and-inflect": translate into uninflected words and predict inflection based on source side

Chahuneau et al. (2013)

Word Segmentation

Chinese Word Segmentation

- Word segmentation: some languages including Chinese are totally untokenized
- LSTMs over character embeddings / character bigram embeddings to predict word boundaries
- Having the right segmentation can help machine translation

冬 天 (winter)， 能 (can) 穿 (wear) 多 少 (amount) 穿 (wear) 多 少 (amount)； 夏 天 (summer)， 能 (can) 穿 (wear) 多 (more) 少 (little) 穿 (wear) 多 (more) 少 (little)。

Without the word "夏天 (summer)" or "冬天 (winter)", it is difficult to segment the phrase "能穿多少穿多少".

- separating nouns and pre-modifying adjectives:
  高血压 (*high blood pressure*)
  → 高(*high*) 血压(*blood pressure*)
- separating compound nouns:
  内政部 (*Department of Internal Affairs*)
  → 内政(*Internal Affairs*) 部(*Department*).

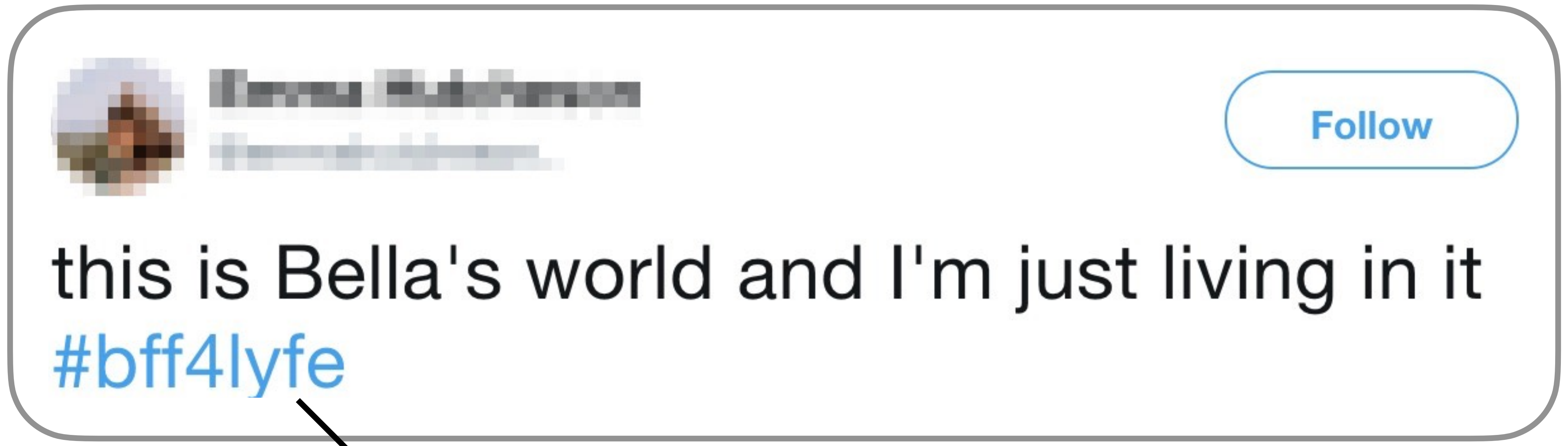
Chen et al. (2015)

English Word Segmentation?

A case study: Hashtag Segmentation

income inequality # debate night

conveys the **topic** of the tweet



bff 4 lyfe

conveys the **sentiment** of the tweet

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Challenges: entities, abbreviations, non-standard spellings, slang …

pawpawty
Microsoft's WordBreaker
(Wang et al., 2011)

pawpaw ty
GATE's hashtag tokenizer
(Maynard & Greenword, 2014)

pawpawty
(Çelebi & Özgür, 2017)

paw pawty
HashtagMaster
(Our Work) ✓

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

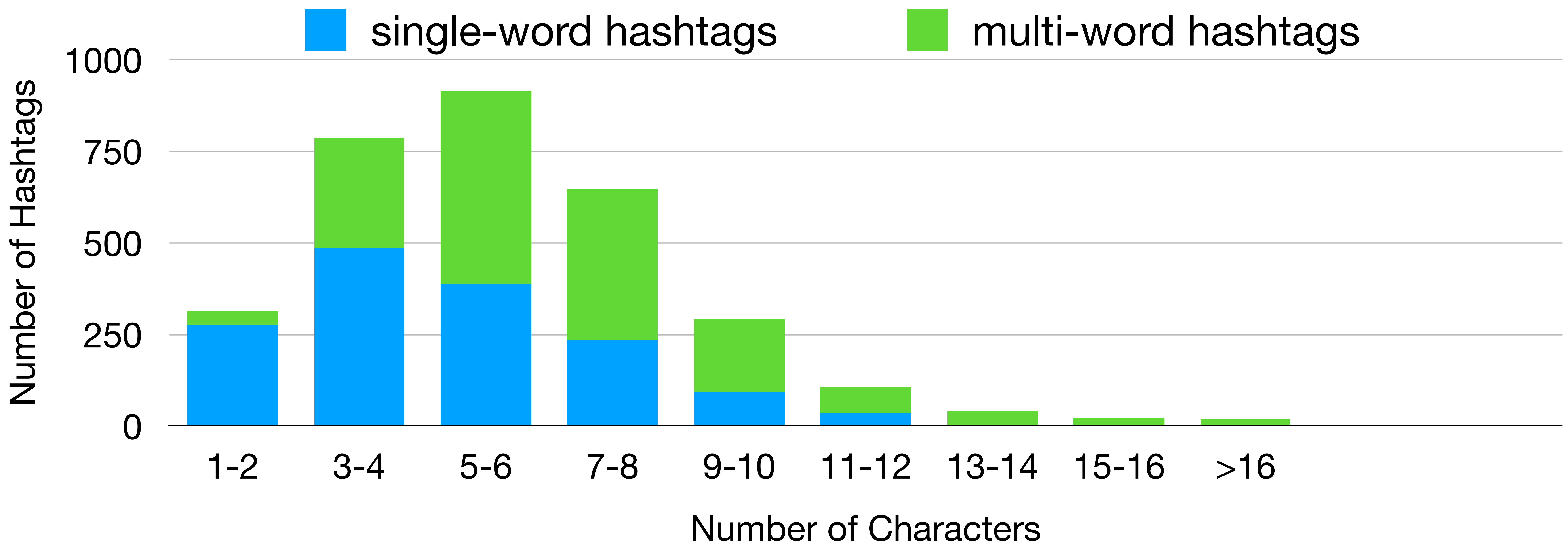
- N-gram language models trained on Twitter data can rank candidate segmentations pretty well. **But,** smoothing is tricky …

	ngram LM (Kneser-Ney)	ngram LM (Good-Turing)	Linguistic Features
#mamapedia → mamapedia #foodstagram → foodstagram	✓	✗	✗
#winebarsf → wine bar sf #wewantmcfly → we want mcfly	✗	✓	✗
#TechLunchSouth → Tech Lunch South #tinthepark → t in the park	✗	✗	✓

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Most hashtags have <15 characters. We can (almost) enumerate all 2^(1-len) possible segmentations.



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- It’s also very hard to tell apart the top-ranked ones.

input hashtag

h: #songsongaddafisitunes

s_1: # song song addafis itunes

s_2: **# songs on gaddafi s itunes**

s_3: # songs on gaddaf is itunes

....

s_k: # song son gaddafis itunes

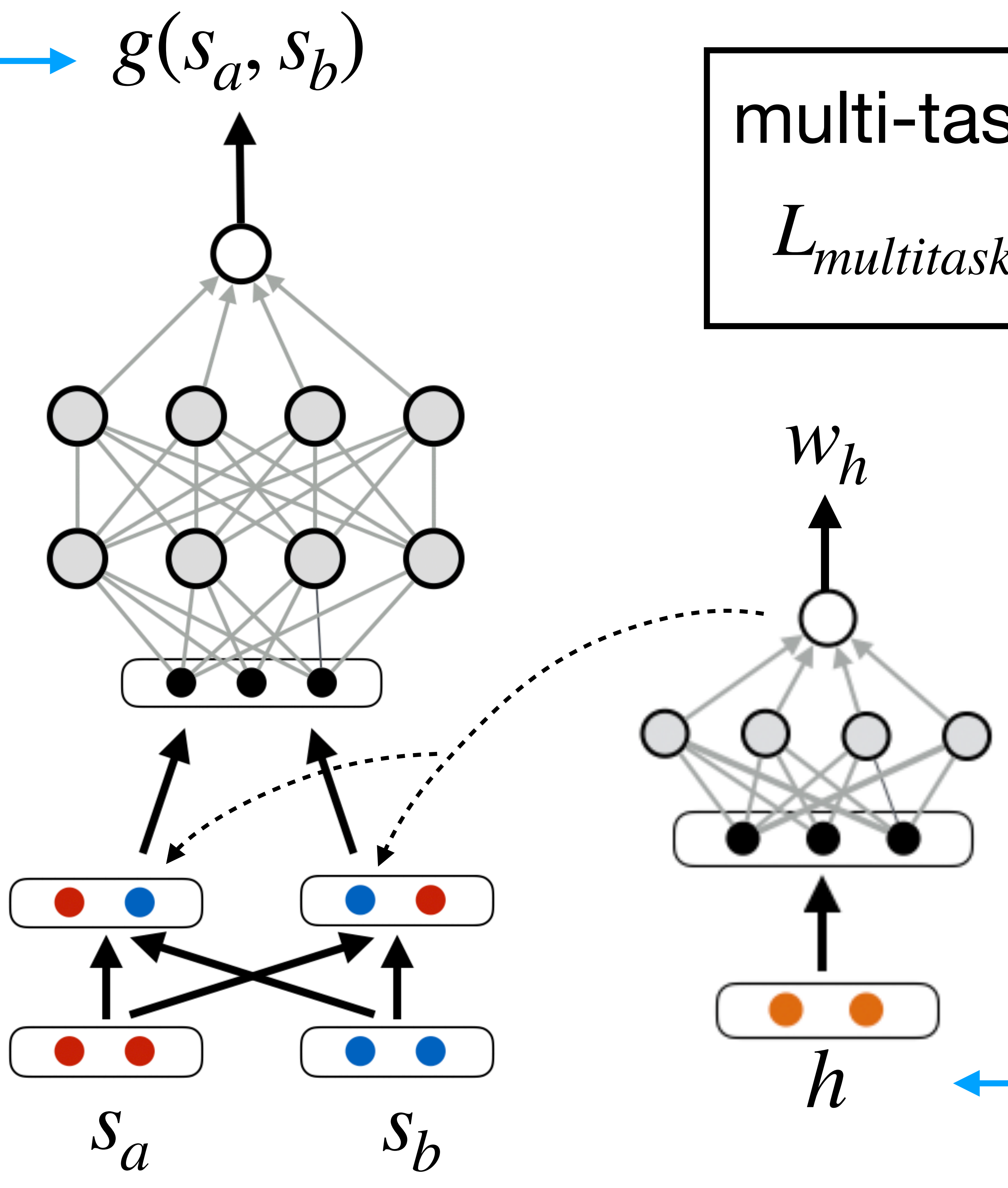
candidate segmentations (top-k)

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Solution: pairwise ranking!

$g(s_a, s_b) > 0$ means s_a is better



multi-task learning objective

$$L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$$

input hashtag

#songsongaddafisitunes

a pair of candidate segmentations

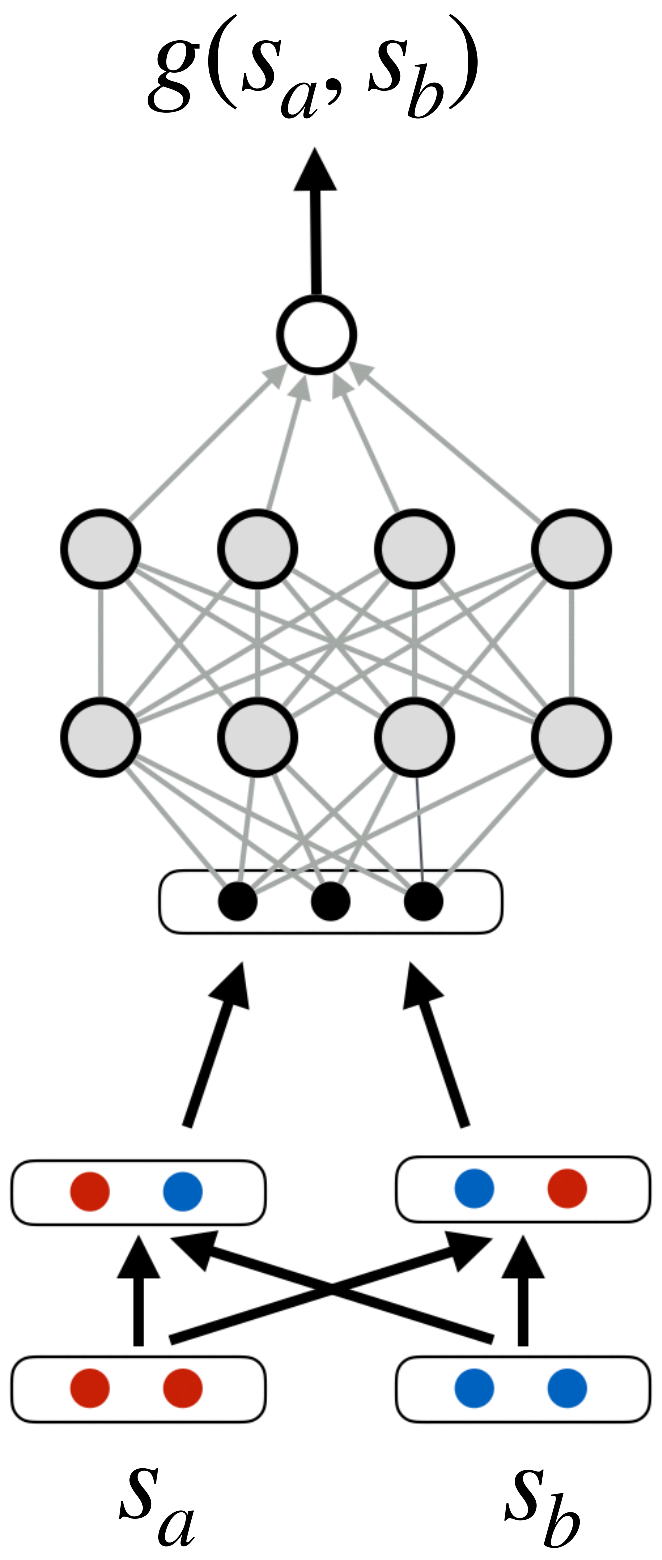
songs on gaddafis itunes

song son gaddafi s itunes

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

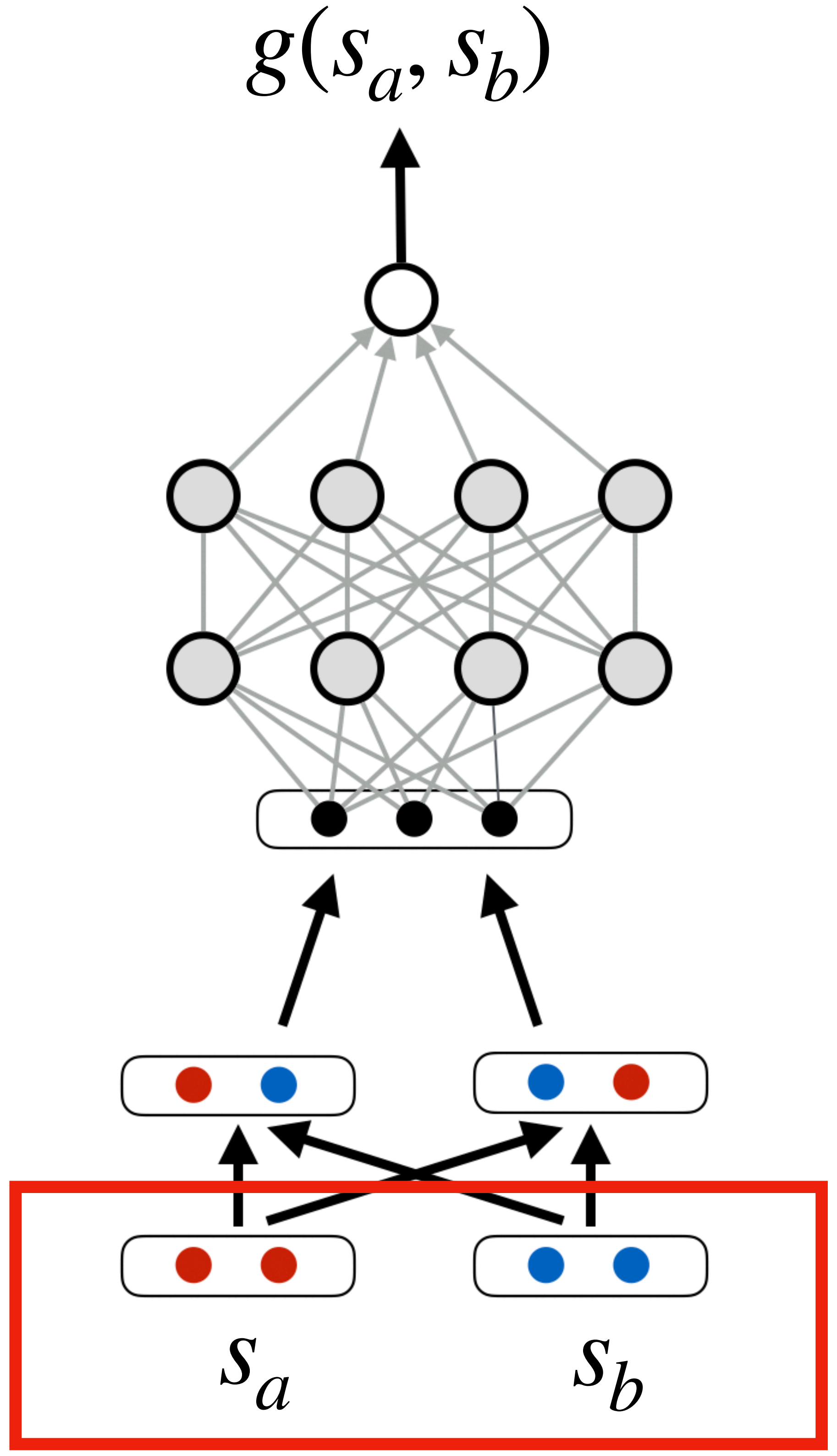
- So we can more easily compare very similar segmentations. We rerank the top-k candidates.



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- The neural pairwise ranking model uses a small number of numerical/ binary features.



Good Turing Smoothing
- Twitter
- Gigaword

Kneser-Ney Smoothing
- Twitter
- Gigaword

Ngram Language Model Probabilities

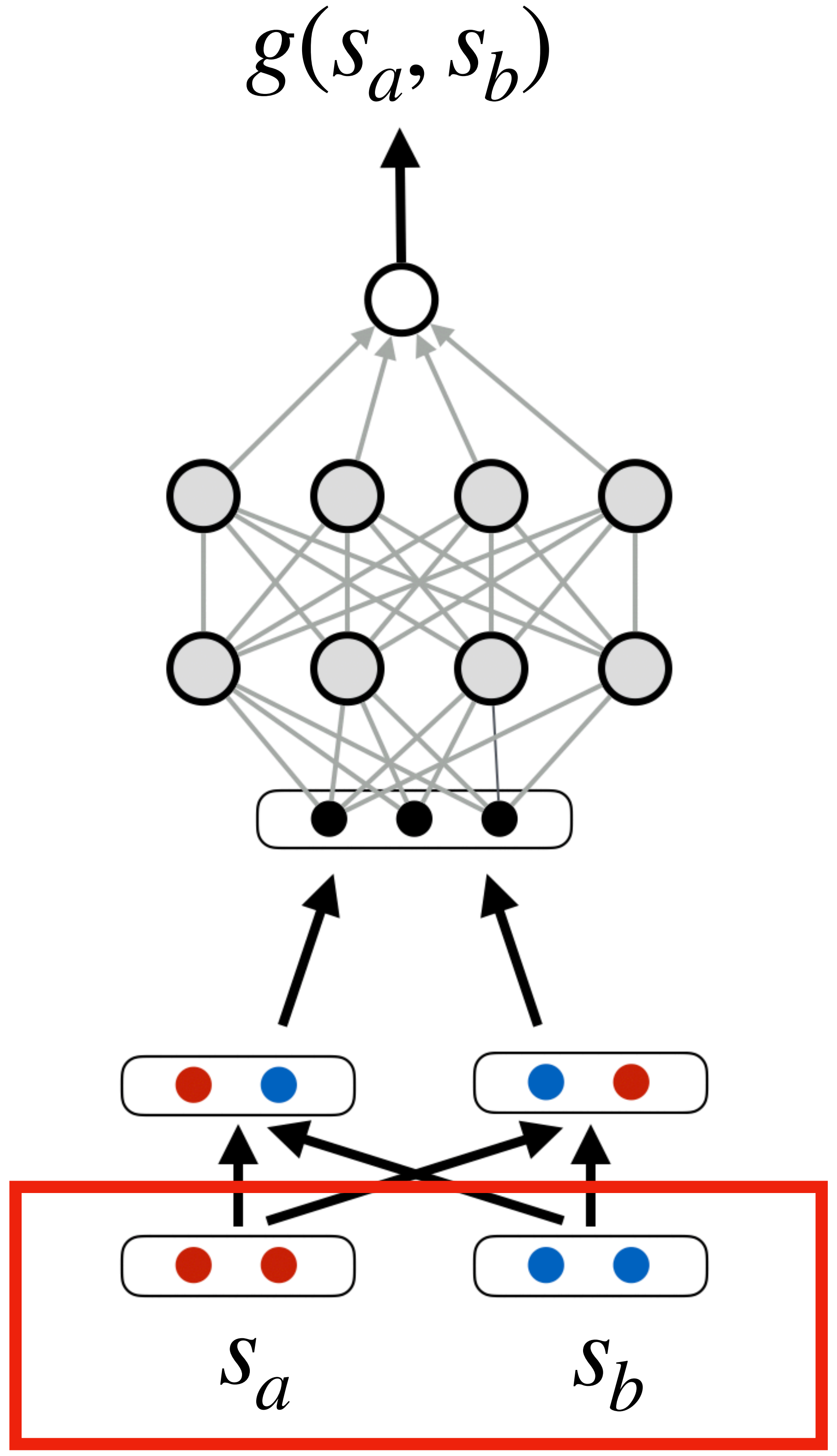
Word length
Number of words
Word shapes
Urban Dictionary
Named entities
Google counts

Linguistic Features

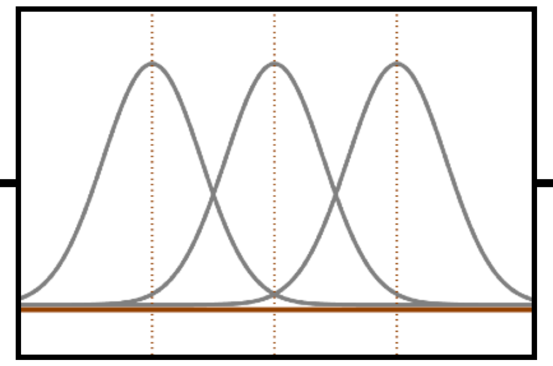
Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Vectorize numerical/binary features.



Gaussian Vectorization

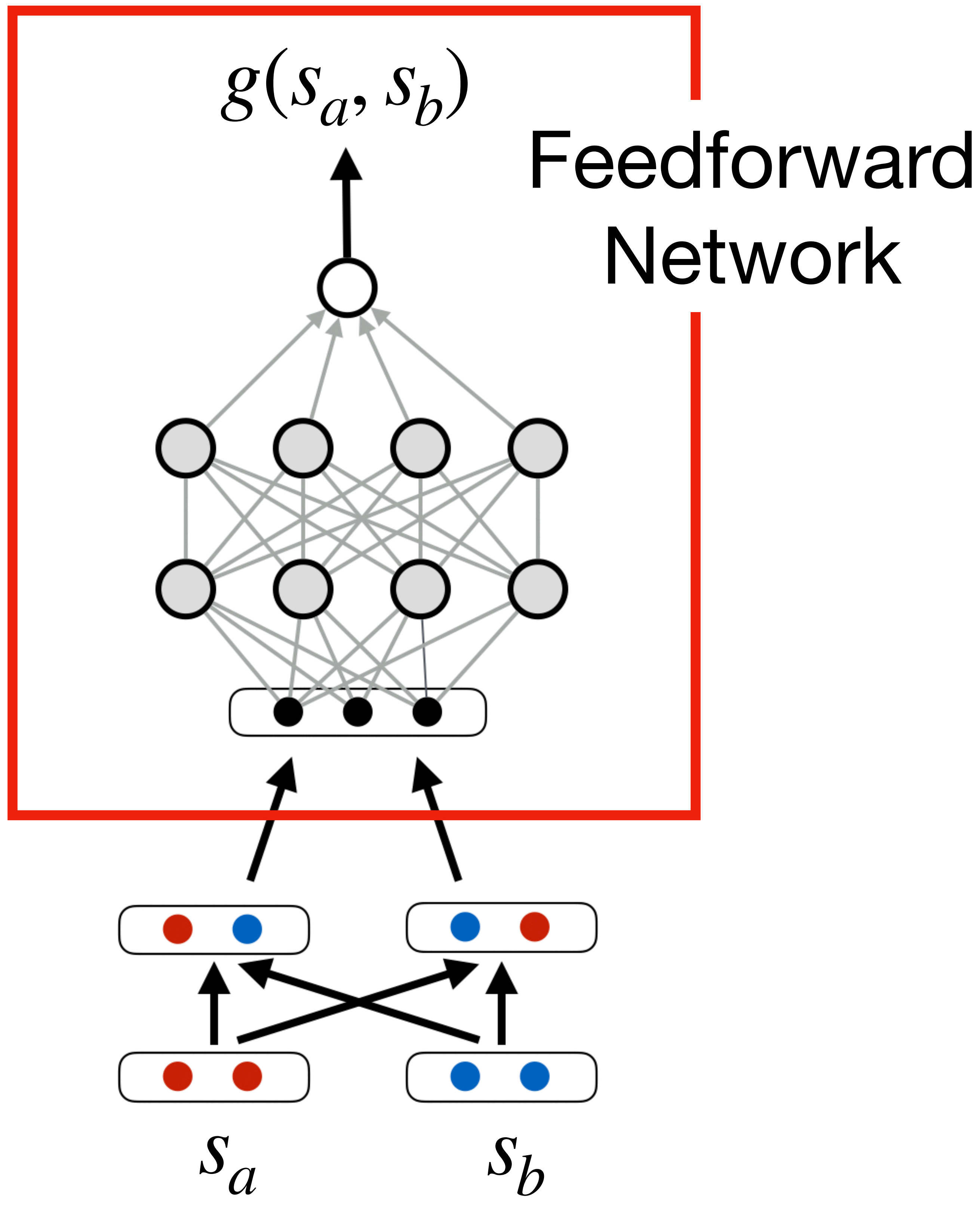
$f_1(s_a) = 0.41$ →  → $\overrightarrow{f_1(s_a)}$ = [~0.0, **0.44, 0.54,** ~0.02, ~0.0]

$$d_j(f(\cdot)) = e^{-\frac{(f(\cdot)-\mu_j)^2}{2\sigma^2}}$$

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Trained with mean squared error (MSE) or margin ranking loss.



Predicted Pairwise Score

$$L_{MSE} = \frac{1}{m} \sum_{i=1}^{m} (g^{*(i)}(s_a, s_b) - g^{(i)}(s_a, s_b))^2$$

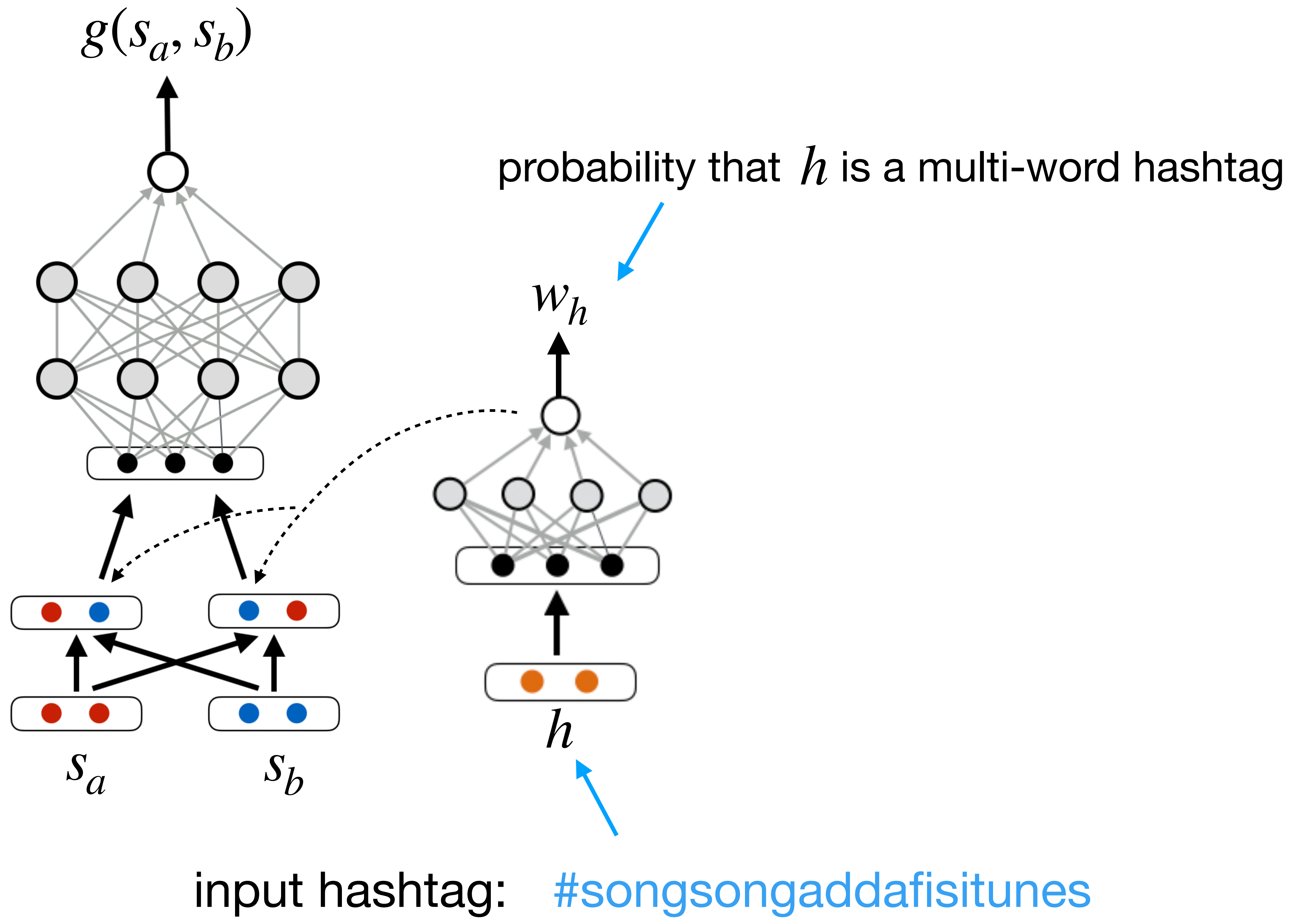
Gold Pairwise Score

$g^*(s_a, s_b) = sim(s_a, s^*) - sim(s_b, s^*)$, where s^* is the gold segmentation.

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

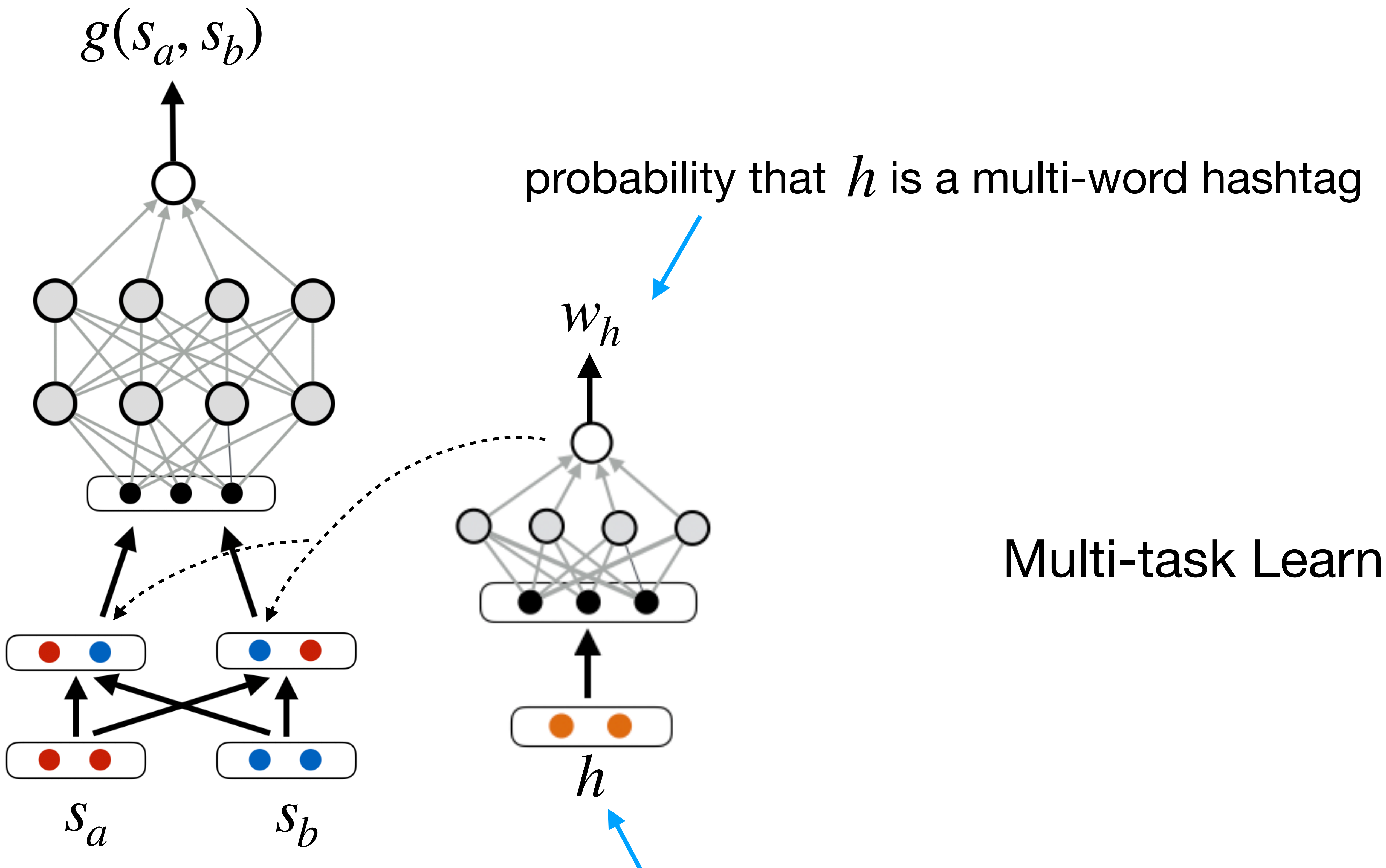
- Adaptive multi-task learning: as different features work for single- vs. multi-word hashtags, we introduce a binary classification task.



Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Adaptive multi-task learning: as different features work for single- vs. multi-word hashtags, we introduce a binary classification task.



Pairwise Ranking (main)

Multi-task Learning Objective: $L_{multitask} = \lambda_1 L_{MSE} + \lambda_2 L_{BCE}$

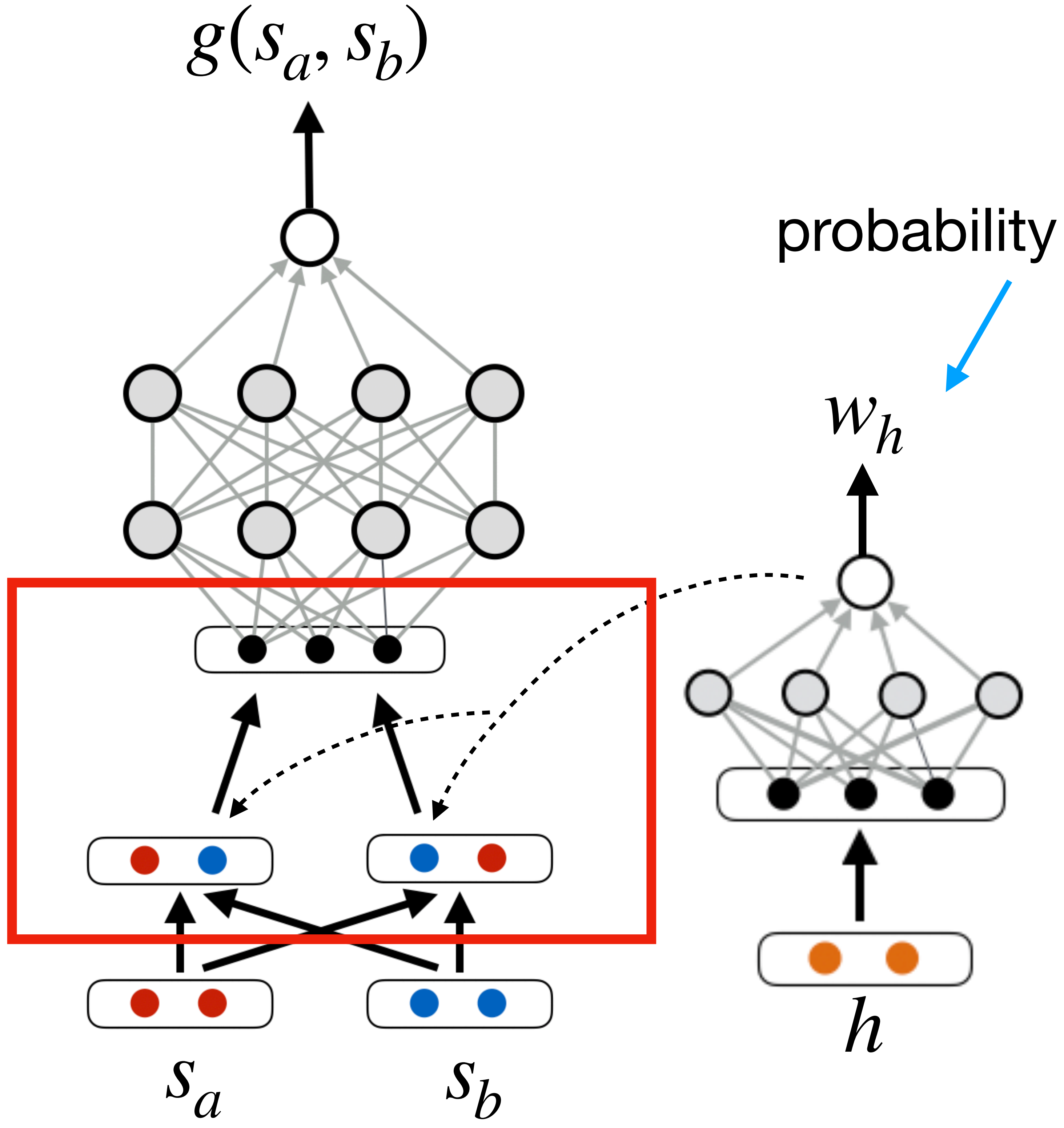
Binary Classification (auxiliary)

input hashtag: #songsongaddafisitunes

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. “Multi-task Pairwise Neural Ranking for Hashtag Segmentation” in ACL (2019)

Hashtag Segmentation

- Adaptive multi-task learning: as different features work for single- vs. multi-word hashtags, we introduce a binary classification task.



probability that h is a multi-word hashtag

Good-Turing + Linguistic Features **Kneser-Ney**

Adaptive Gating: $g(s_a, s_b) = FFNN\left(w_h \mathbf{s}_{ab}^{GL} + (1 - w_h)\mathbf{s}_{ab}^{KN}\right)$

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Hashtag Segmentation

- Error Analysis: some hashtags are just hard … our model almost gets them right (Accuracy@2 is ~98%).

Rare Words	#OHIOis4thrillaz	→	OHIO is 4th rillaz	✗
			OHIO is 4 thrillaz	✓
Abbreviations	#BTVSMB	→	BTVSMB	✗
			BTV SMB (Burlington VT Social Media Breakfast)	✓
Misspellings	#wolframapltha	→	wolfram apltha	✗
			wolframapltha	✓
Others	#iseelondoniseeparis	→	isee london isee paris	✗
			I see london i see paris	✓

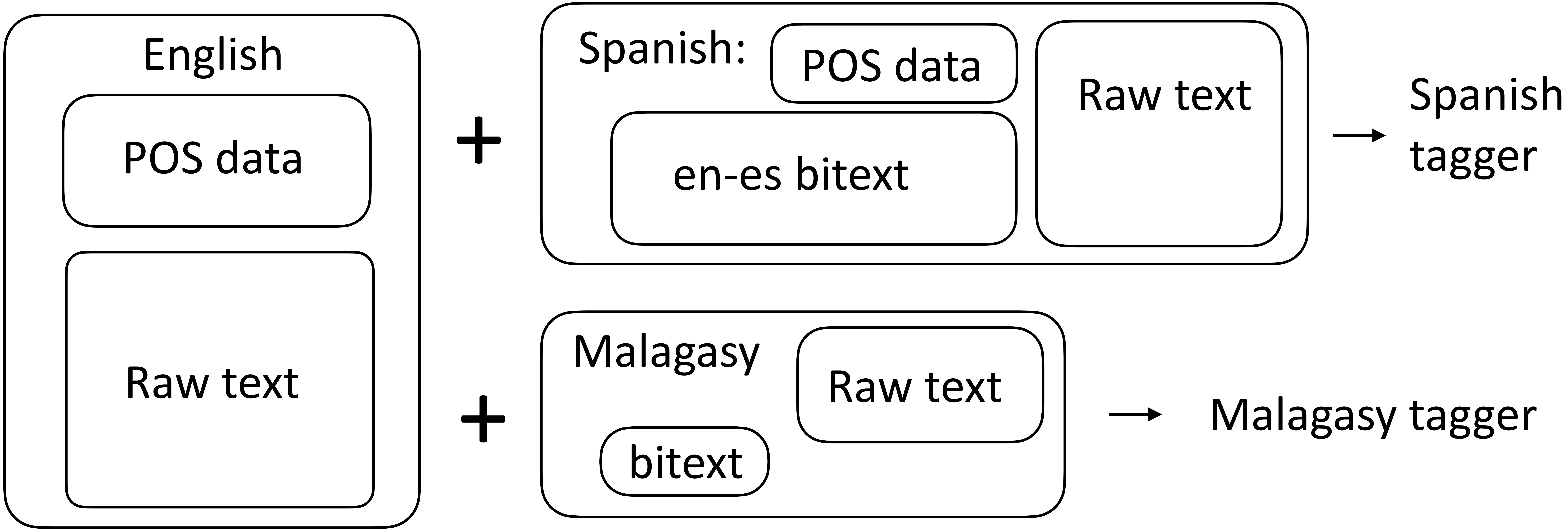
WolframAlpha

Mounica Maddela, Wei Xu, Daniel Preotiuc-Pietro. "Multi-task Pairwise Neural Ranking for Hashtag Segmentation" in ACL (2019)

Cross-Lingual Tagging and Parsing

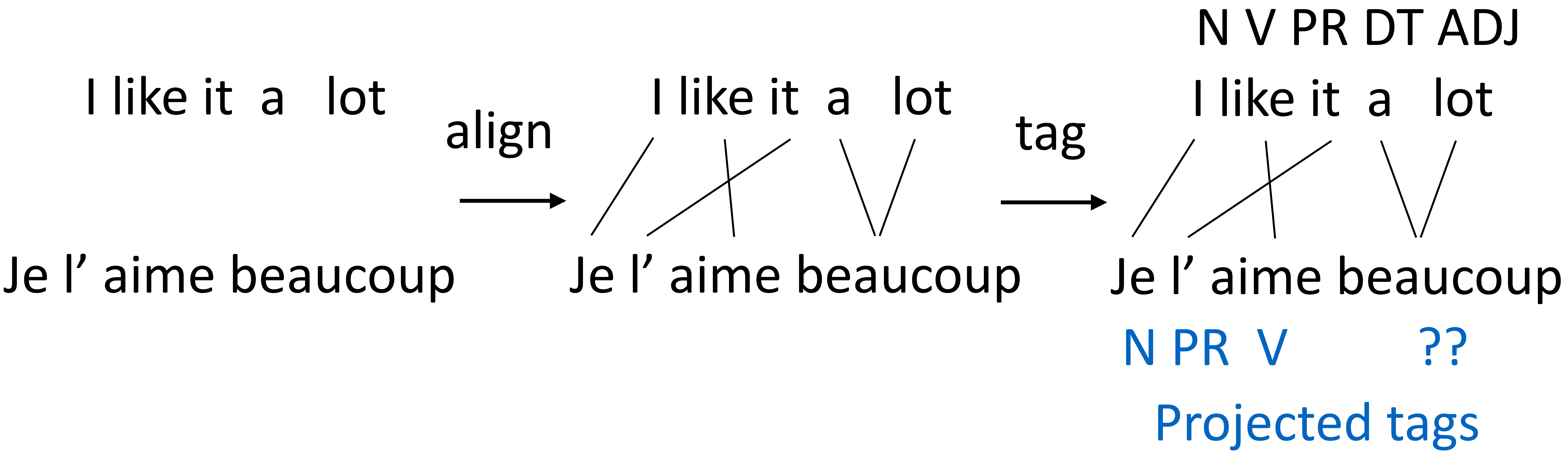
Cross-Lingual Tagging

- Labeling POS datasets is expensive
- Can we transfer annotation from *high-resource* languages (English, etc.) to *low-resource* languages?



Cross-Lingual Tagging

- Can we leverage word alignment here?

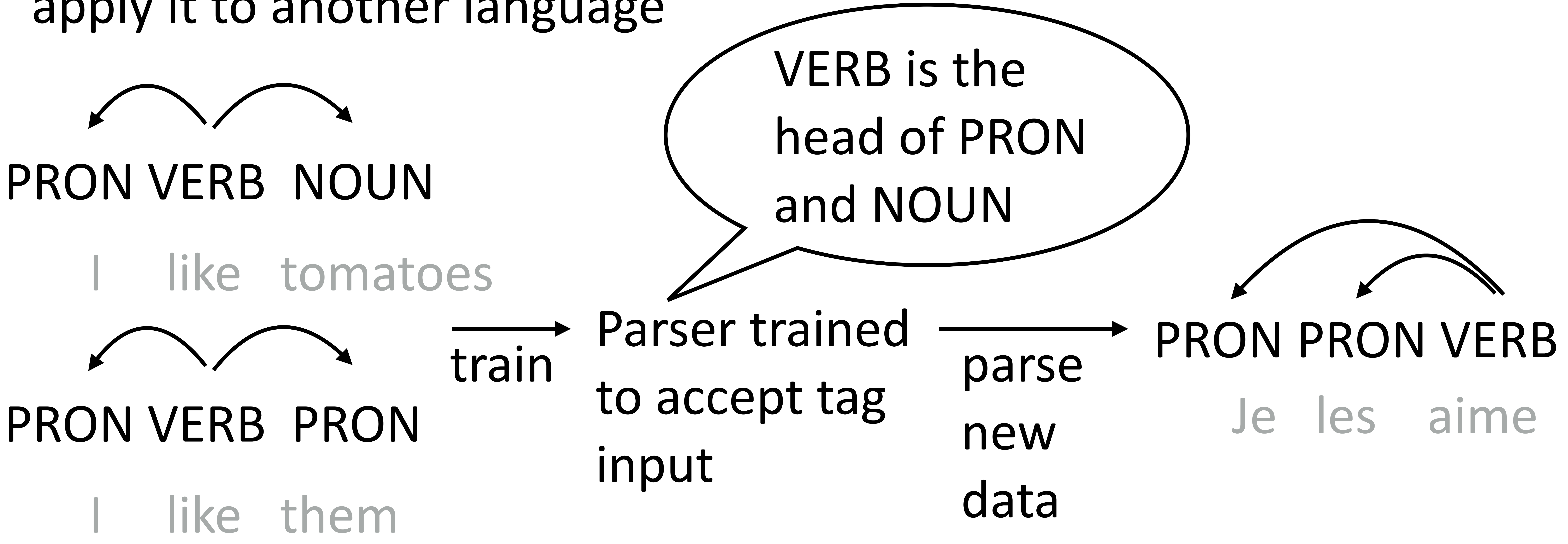


- Tag with English tagger, project across bitext, train French tagger? Works pretty well

Das and Petrov (2011)

Cross-Lingual Parsing

- Now that we can POS tag other languages, can we parse them too?
- Direct transfer: train a parser over POS sequences in one language, then apply it to another language



McDonald et al. (2011)

Cross-Lingual Parsing

	best-source		avg-source	gold-POS		pred-POS	
	source	gold-POS	gold-POS	multi-dir.	multi-proj.	multi-dir.	multi-proj.
da	it	48.6	46.3	48.9	49.5	46.2	47.5
de	nl	55.8	48.9	56.7	56.6	51.7	52.0
el	en	63.9	51.7	60.1	65.1	58.5	63.0
es	it	68.4	53.2	64.2	64.5	55.6	56.5
it	pt	69.1	58.5	64.1	65.0	56.8	58.9
nl	el	62.1	49.9	55.8	65.7	54.3	64.4
pt	it	74.8	61.6	74.0	75.6	67.7	70.3
sv	pt	66.8	54.8	65.3	68.0	58.3	62.1
avg		63.7	51.6	61.1	63.8	56.1	59.3

- Multi-dir: transfer a parser trained on several source treebanks to the target language
- Multi-proj: more complex annotation projection approach

McDonald et al. (2011)

Cross-Lingual Word Representations

Multilingual Embeddings

- Input: corpora in many languages. Output: embeddings where similar words *in different languages* have similar embeddings

I have an apple
47 24 18 427

J’ ai des oranges
47 24 89 1981

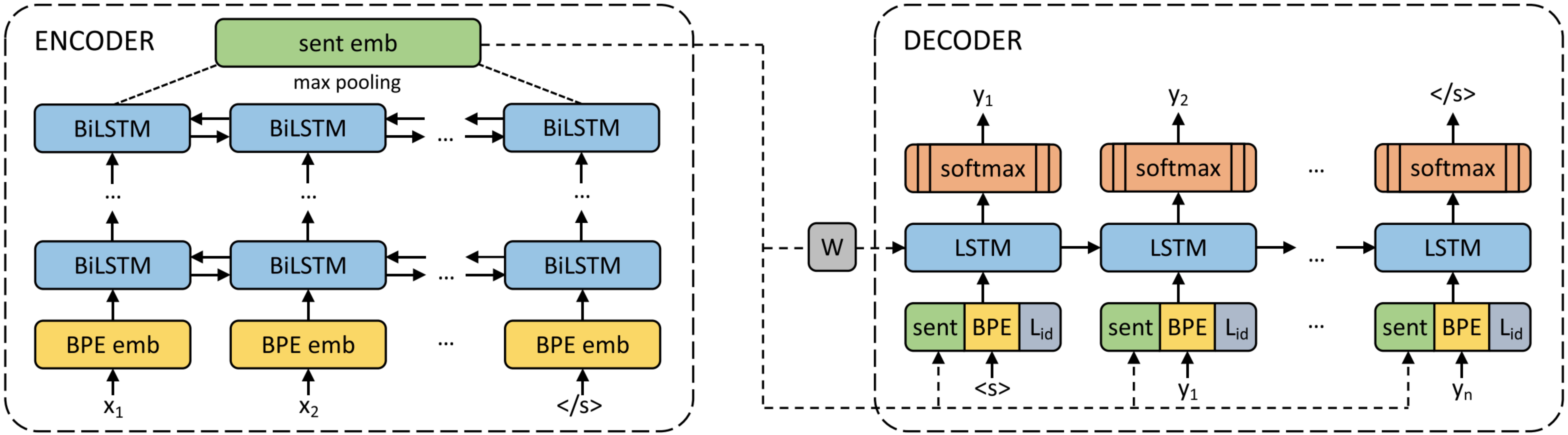
ID: 24
ai have

ID: 47
I Je J’

- multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train “monolingual” embeddings over all these corpora
- Works okay but not all that well

Ammar et al. (2016)

Multilingual Sentence Embeddings



- Form BPE vocabulary over all corpora (50k merges); will include characters from every script
- Take a bunch of bitexts and train an MT model between a bunch of language pairs with shared parameters, use W as sentence embeddings

Artetxe et al. (2019)

Multilingual Sentence Embeddings

		EN	EN → XX													
			fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Zero-Shot Transfer, one NLI system for all languages:																
Conneau et al.	X-BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
(2018b)	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2
BERT uncased*	Transformer	<u>81.4</u>	–	<u>74.3</u>	70.5	–	–	–	–	62.1	–	–	63.8	–	–	58.3
Proposed method	BiLSTM	73.9	**71.9**	72.9	<u>72.6</u>	**72.8**	**74.2**	**72.1**	**69.7**	**71.4**	**72.0**	**69.2**	<u>71.4</u>	**65.5**	**62.2**	<u>61.0</u>

- Train a system for NLI (entailment/neutral/contradiction of a sentence pair) on English and evaluate on other languages

Artetxe et al. (2019)

Multilingual BERT

- Take top 104 Wikipedias, train BERT on all of them simultaneously
- What does this look like?

Beethoven may have proposed unsuccessfully to Therese Malfatti, the supposed dedicatee of "Für Elise"; his status as a commoner may again have interfered with those plans.

当人们在马尔法蒂身后发现这部小曲的手稿时，便误认为上面写的是“Für Elise”（即《给爱丽丝》）[51]。

Кита́й (официально — Кита́йская Наро́дная Респу́блика, сокращённо — КНР; кит. трад. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэньминь Гунхэго) — государство в Восточной Аз

Devlin et al. (2019)

Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

- Can transfer BERT directly across languages with some success
- …but this evaluation is on languages that all share an alphabet

Pires et al. (2019)

Multilingual BERT: Results

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- Urdu (Arabic/Nastaliq script) => Hindi (Devanagari). Transfers well despite different alphabets!

- Japanese => English: different script and very different syntax

Pires et al. (2019)

Scaling Up: XLM-RoBERTa (XLM-R)

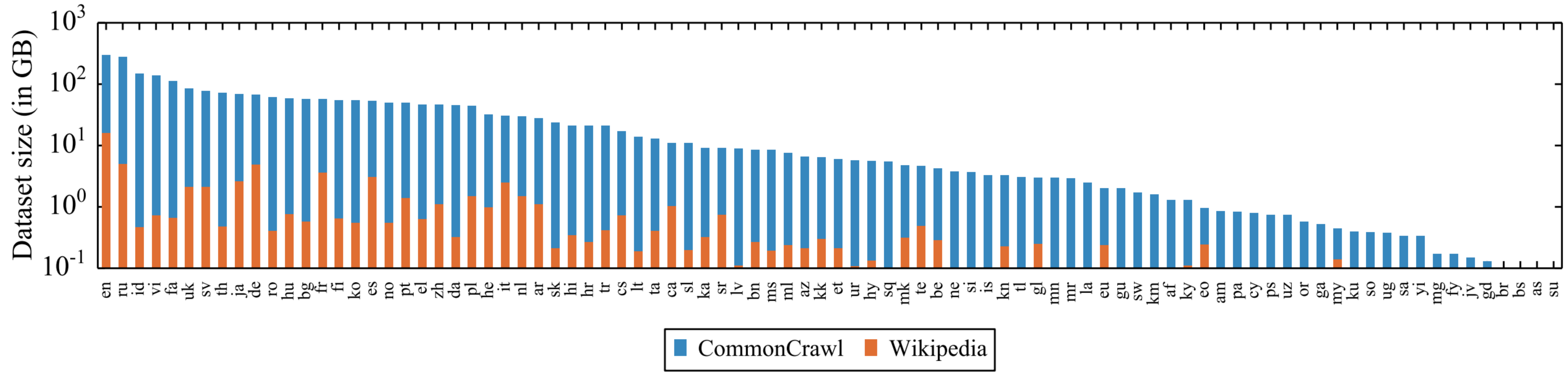


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- Larger "Common Crawl" dataset, better performance than mBERT
- Low-resource languages benefit from training on other languages
- High-resource languages see a small performance hit, but not much

Conneau et al. (2019)

Scaling Up: Benchmarks

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction
	MLQA			4,517–11,590	translations	7	Span extraction
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval

- Many of these datasets are translations of base datasets, not originally annotated in those languages
- Exceptions: POS, NER, TyDiQA

Hu et al. (2021)

TyDiQA

- Typologically-diverse QA dataset
- Annotators write questions based on very short snippets of articles; answers may or may not exist, fetched from elsewhere in Wikipedia

Language	**Train** *(1-way)*	**Dev** *(3-way)*	**Test** *(3-way)*
(English)	9,211	1031	1046
Arabic	23,092	1380	1421
Bengali	10,768	328	334
Finnish	15,285	2082	2065
Indonesian	14,952	1805	1809
Japanese	16,288	1709	1706
Kiswahili	17,613	2288	2278
Korean	10,981	1698	1722
Russian	12,803	1625	1637
Telugu	24,558	2479	2530
Thai	11,365	2245	2203
TOTAL	**166,916**	**18,670**	**18,751**

Clark et al. (2021)

TyDiQA

- Typologically-diverse QA dataset
- Annotators write questions based on very short snippets of articles; answers may or may not exist, fetched from elsewhere in Wikipedia

Q: Как далеко Уран от Землии?
how far Uranus-SG.NOM from Earth-SG.GEN?

Q: Как далеко Уран от Земл-и?
how far Uranus-SG.NOM from Earth-SG.GEN?

How far is Uranus from Earth?

A: Расстояние между Уран-ом и Земл-ёй меняется от 2,6 до 3,15 млрд км...
distance between Uranus-SG.INSTR and Earth-SG.INSTR varies from 2,6 to 3,15 bln km...

The distance between Uranus and Earth fluctuates from 2.6 to 3.15 bln km...

Figure 3: Russian example of morphological variation across question-answer pairs due to the difference in syntactic context: the entities are identical but have different representation, making simple string matching more difficult. The names of the planets are in the subject (Уран, Uranus-NOM) and object of the preposition (от земли, from Earth-GEN) context in the question. The relevant passage with the answer has the names of the planets in a coordinating phrase that is an object of a preposition (между Ураном и Землёй, between Uranus-INSTR and Earth-INSTR). Because the syntactic contexts are different, the names of the planets have different case marking.

Clark et al. (2021)

Where are we now?

- Universal dependencies: treebanks (+ tags) for 100+ languages
- Datasets in other languages are still small, so projection techniques may still help
- More corpora in other languages, less and less reliance on structured tools like parsers, and pretraining on unlabeled data means that performance on other languages is better than ever
- Multilingual models seem to be working better and better — but still many challenges for low-resource settings

Takeaways

- Many languages have richer morphology than English and pose distinct challenges
- Problems: how to analyze rich morphology, how to generate with it
- Can leverage resources for English using bitexts
- Multilingual models can be learned in a bitext-free way and can transfer between languages