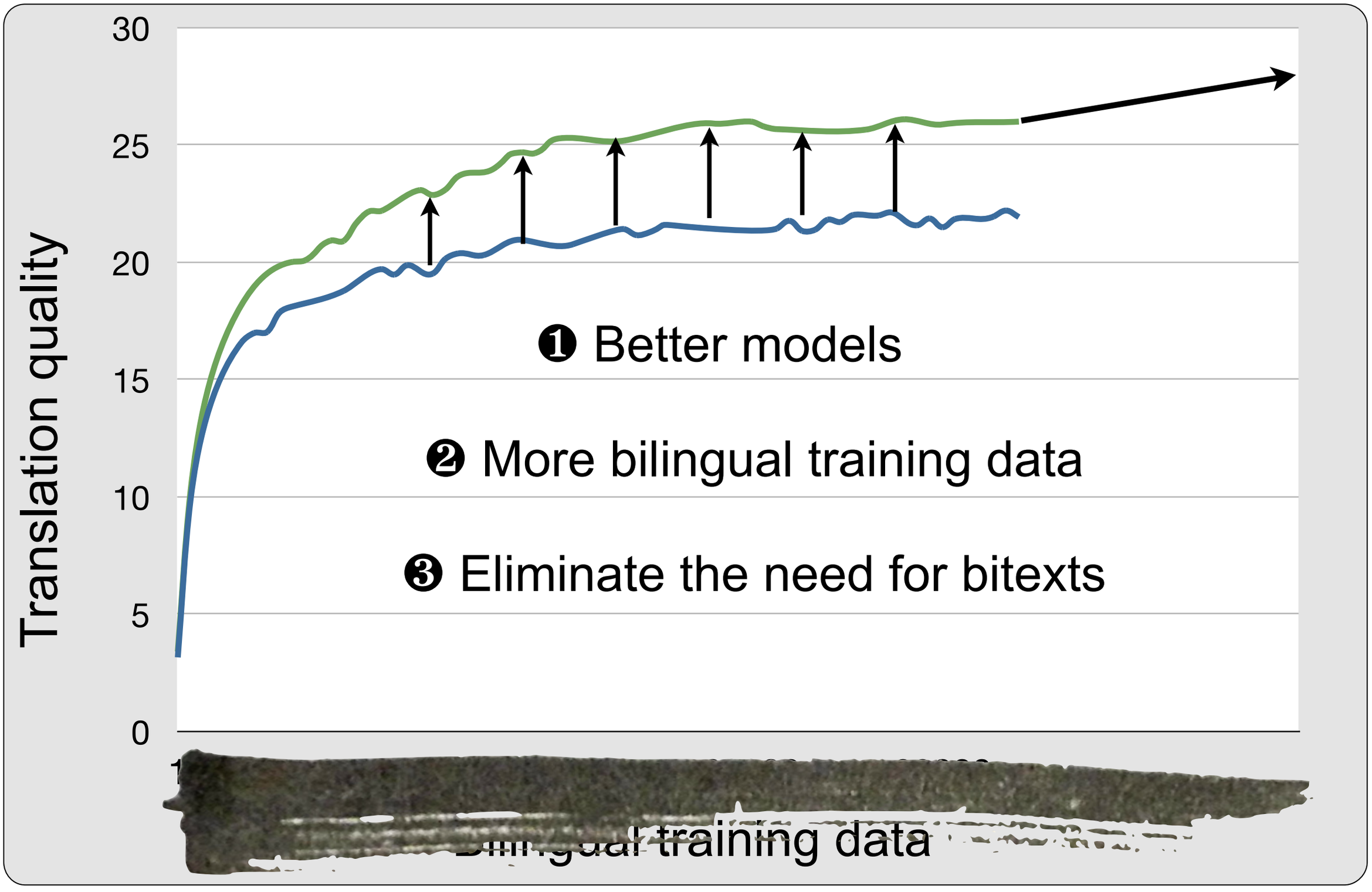
Translation without bilingual parallel corpora

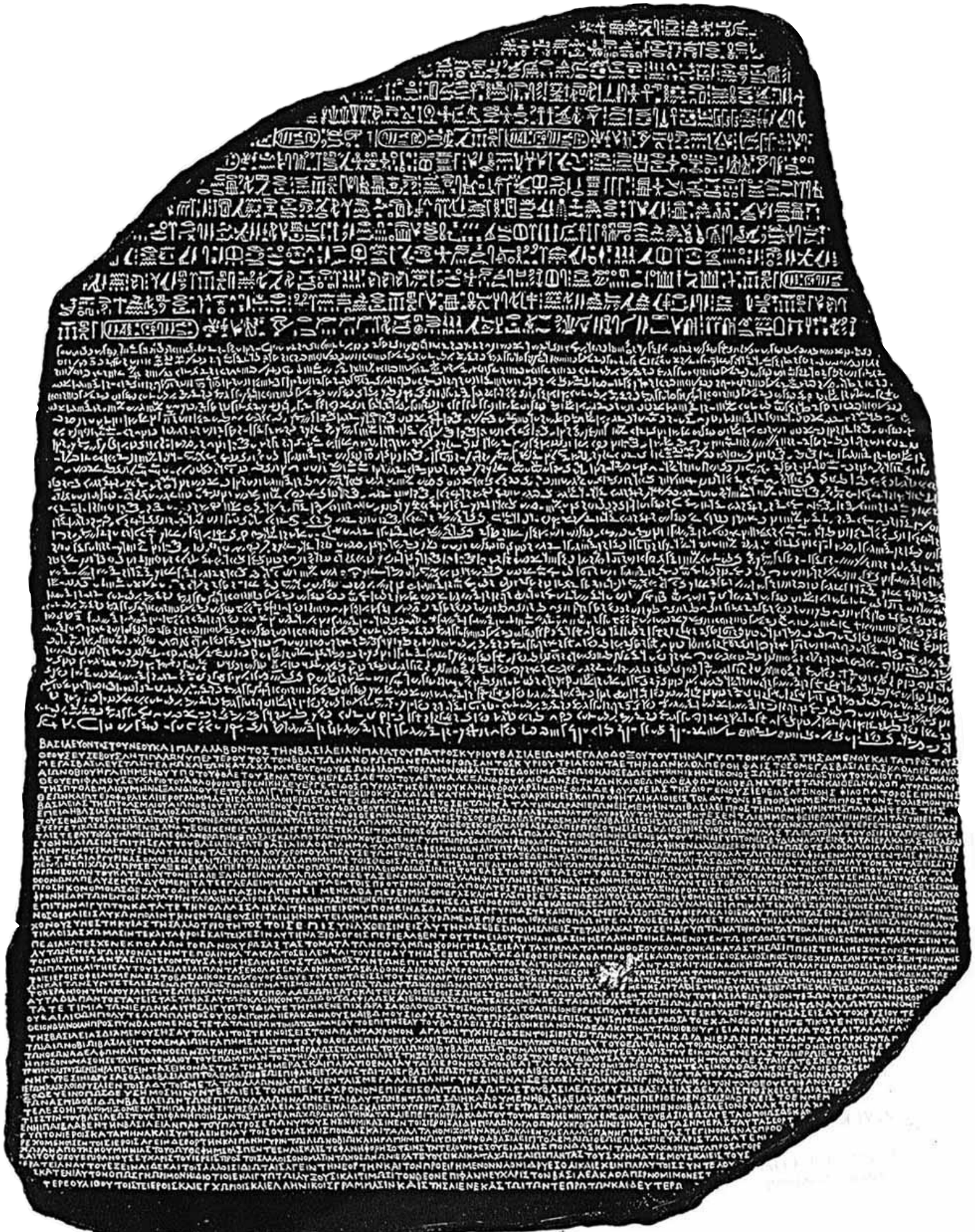
Chris Callison-Burch

Lecture 20

with Ann Irvine, Alex Klementiev, and David Yarowsky

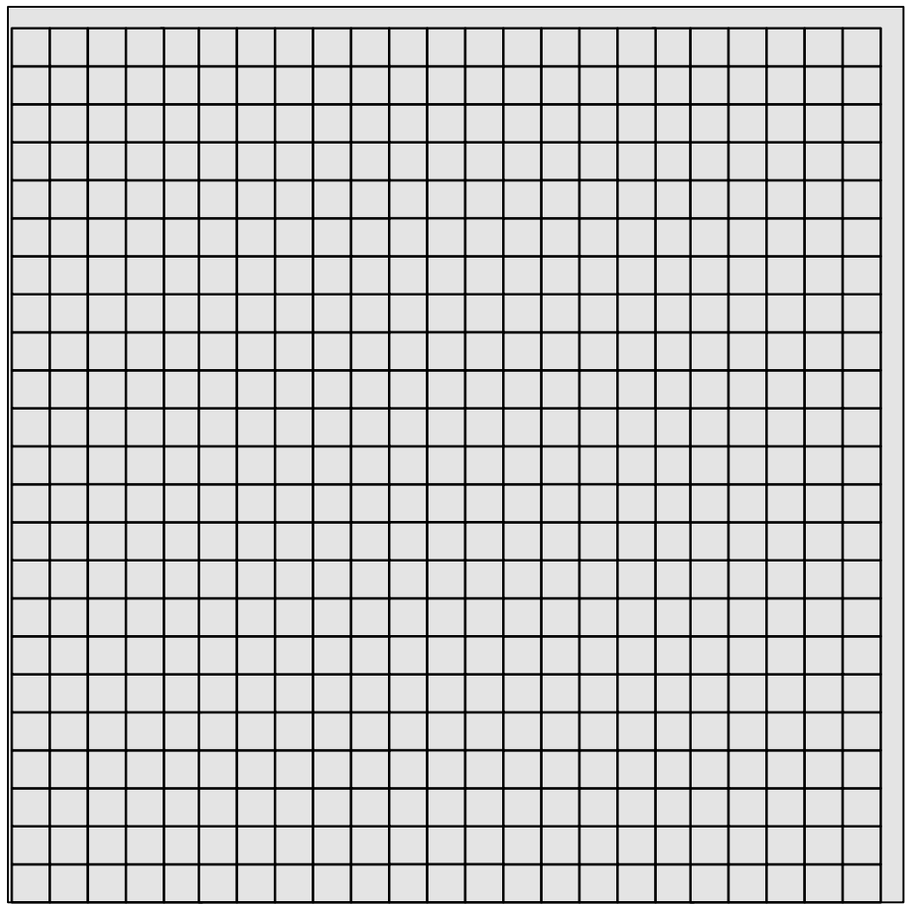
How to Improve Machine Translation





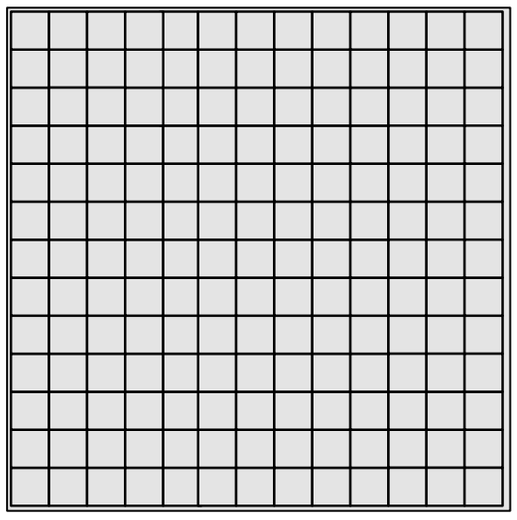
Bilingual data varies by language

1000M



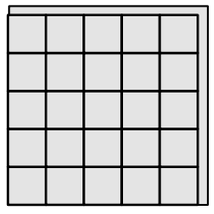
French-English
10^9 word webcrawl

200M



Arabic and Chinese
DARPA GALE

50M



European
Parliament

1.5M

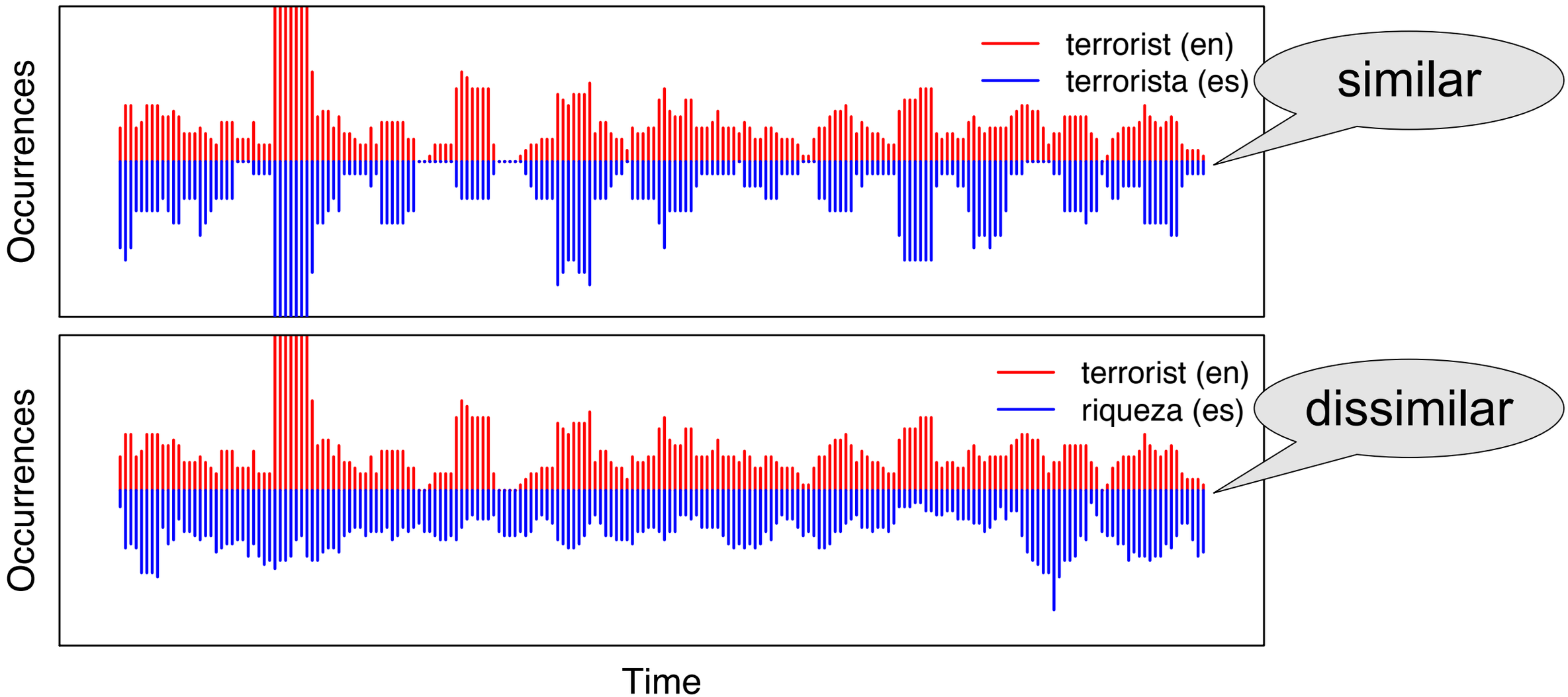
Urdu

Monolingual data is more common

- Typically we have orders of magnitude more monolingual data
- Can we use monolingual data to learn translations?
- Is that a crazy idea?

reclamo otra vez cargos políticos
Fue una demostración de fuerza del aparato gremial. Pidió la reelección de Cristina. Pero insistió en que el sindicalismo tenga candidatos en las listas. Para la CGT, hubo 500 mil personas. Para la Policía y la SIDE, unas diez veces menos.
Cristina saludó por carta y envió a sus ministros
Boda Real
Como una película, que vieron dos mil millones
נשיקה
東亞日報
軍 "北 이상징후 포착" 경계강화
제3노총 6월에 출범
नवदुनिया
1900
200
30
4
प्रधानमंत्री को घेरा
Las Últimas Noticias
tiroteo en que murió Osama bin Laden
מעריב
בן מוות
중앙일보
9·11 10년 '빈 라덴 추격' 끝내다
বিন লাদেন
RIOdePERNAMBUCO
1955 2011
STEVE JOBS
O HOMEM QUE DEU ROSTO AO FUTURO
הארץ
האיש ששינה את חיינו
朝鮮日報
"70兆 날릴 각오" 삼성, 애플과 끝장승부
월街 시위에 美노조 가세
जॉब्स दुनिया से अलविदा

Scoring Translations: Time

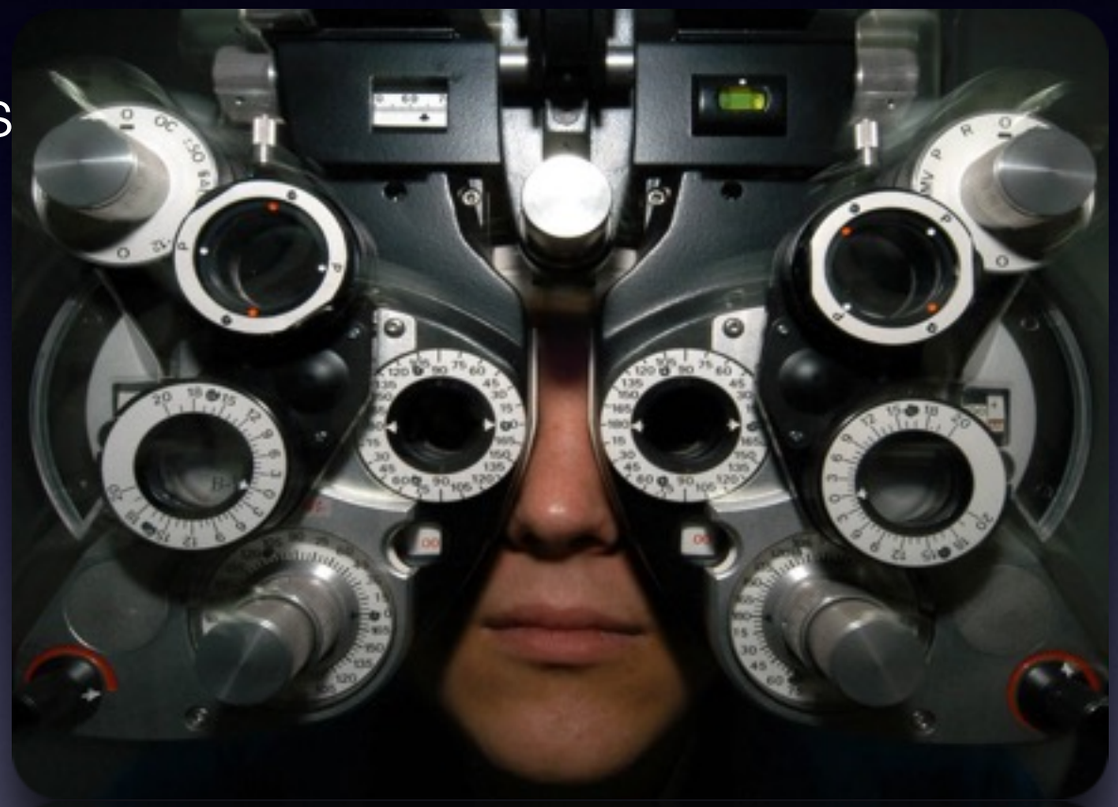


Scoring Translations: Time

eólica	estambul	terrorista	vacuno
wind	**istanbul**	**terrorist**	beef
renewable	erdogan	terrorism	cattle
solar	turkish	terrorists	bse
sources	turkey	attacks	compulsory
renewables	turks	fight	meat
energy	ankara	attack	cows
energies	membership	terror	veal
electricity	negotiations	acts	**cow**
photovoltaic	undcp	threat	labelling
grid	talks	september	papayannakis

Distributional Hypothesis

If we consider **oculist** and **eye-doctor** we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which **oculist** occurs but **lawyer** does not...



It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for **oculist** (not asking what words have the same meaning).

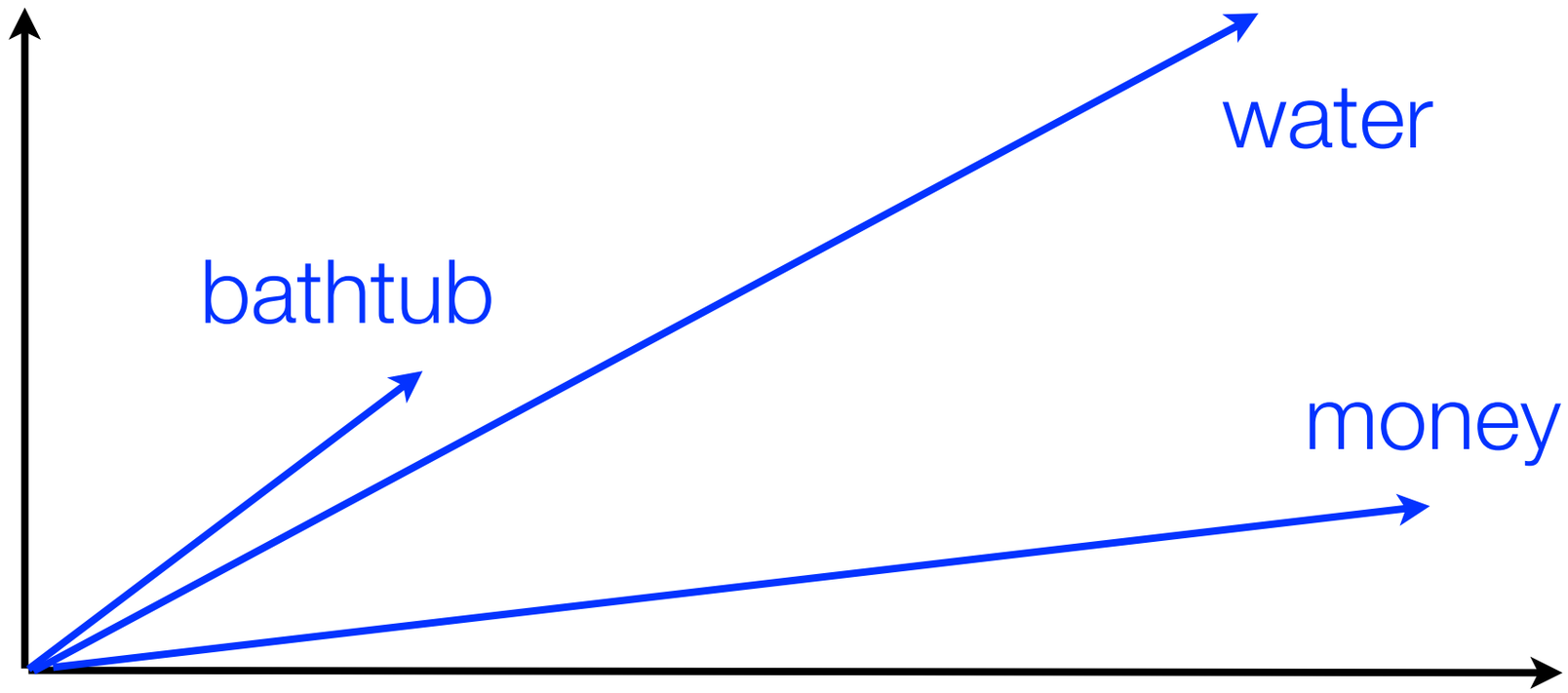
These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.

–Zellig Harris (1954)

Vector Space Models of Word Similarity

Represent a word through the contexts that it has been observed in

He found five fish swimming in an old bathtub.
He slipped down in the bathtub.



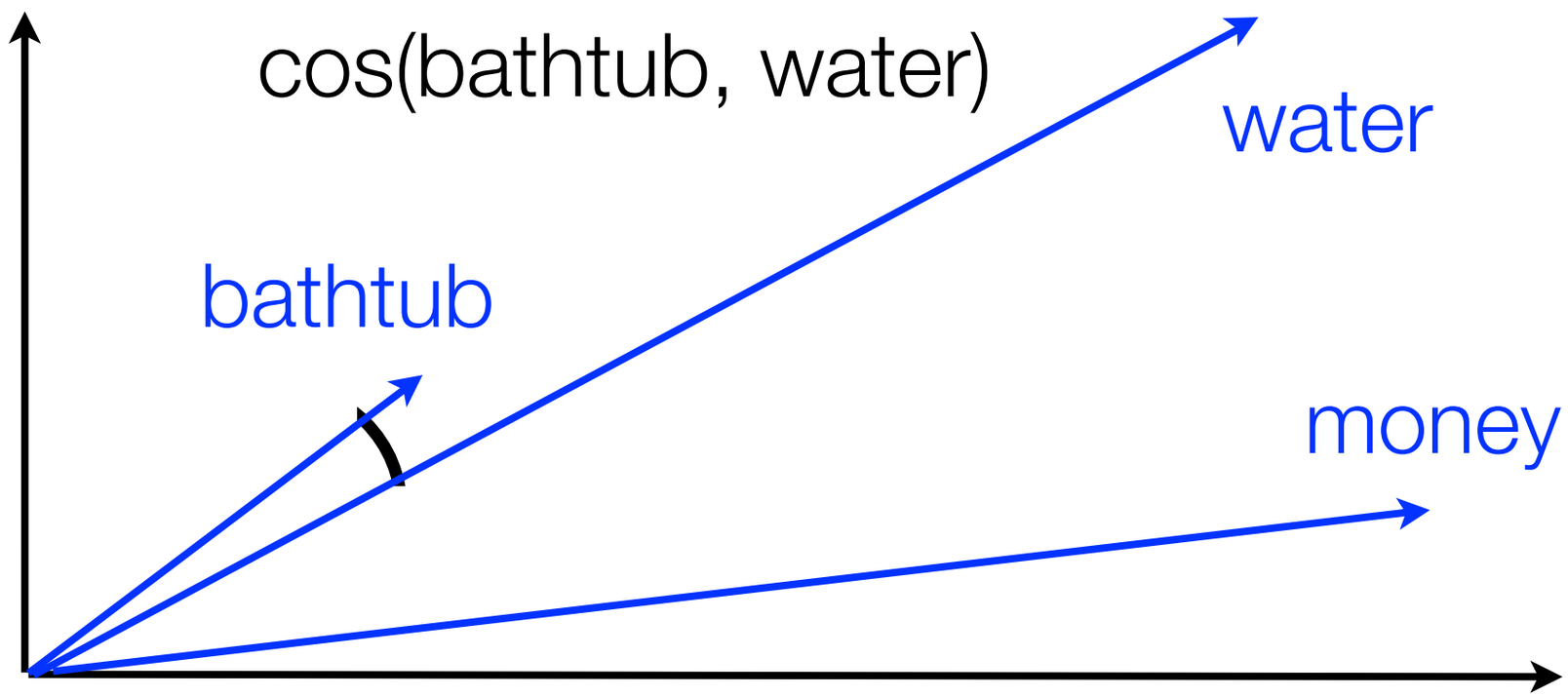
a	1
down	1
find	1
fish	1
five	1
he	2
in	2
slip	1
swim	1
the	1

Vector Space Models of Word Similarity

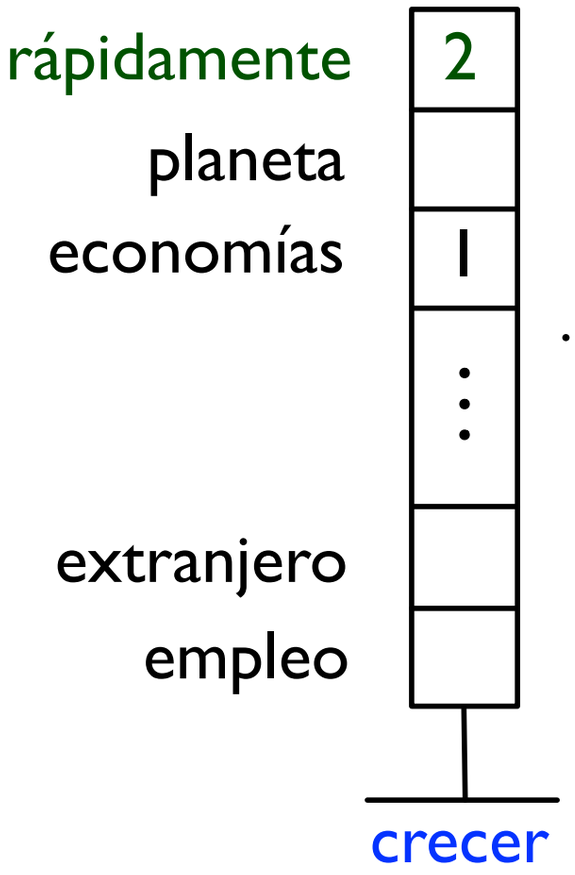
Represent a word through the contexts that it has been observed in

He found five fish swimming in an old bathtub.
He slipped down in the bathtub.

a	1
down	1
find	1
fish	1
five	1
he	2
in	2
slip	1
swim	1
the	1



Scoring Translations: Context

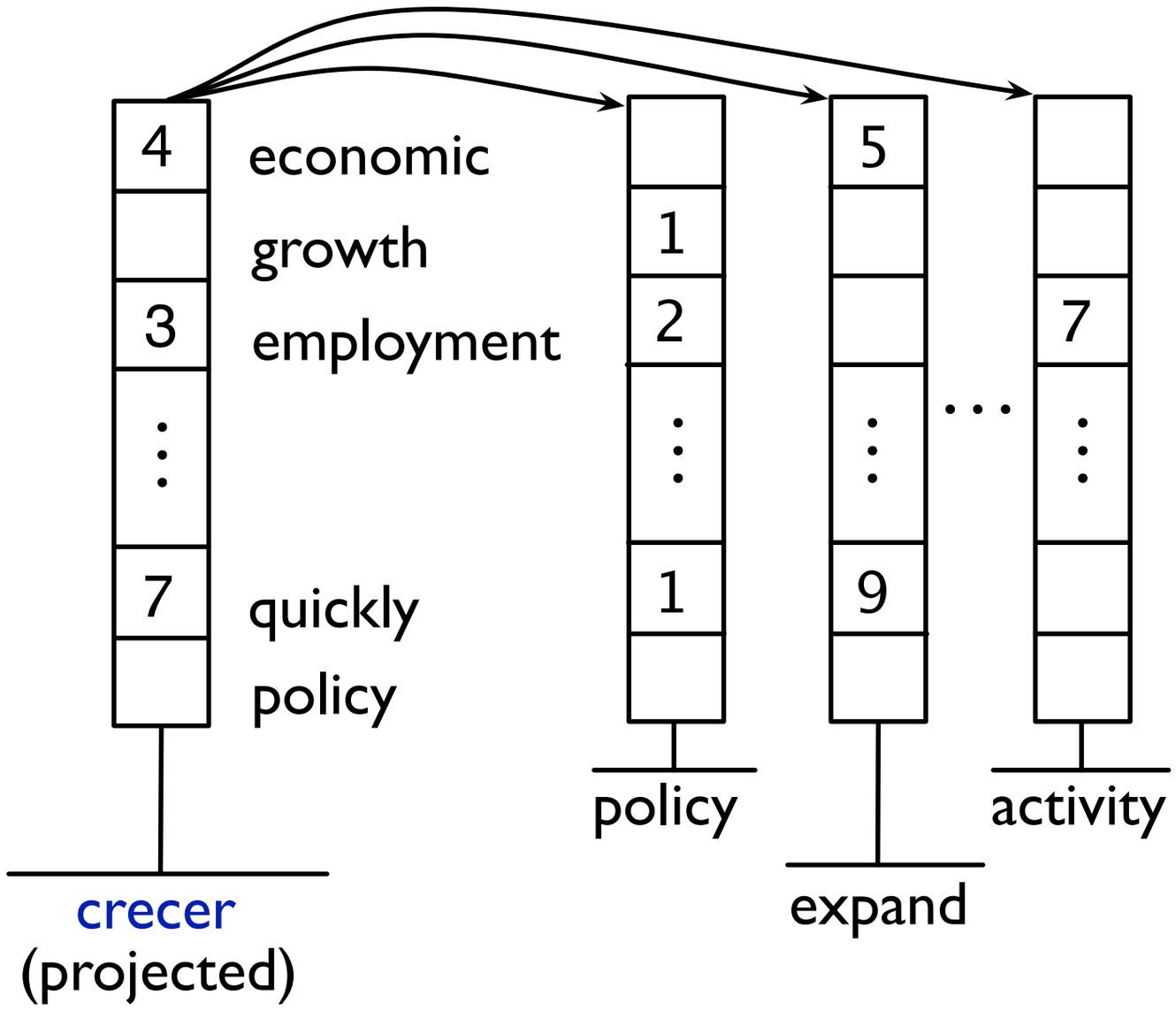


... este número podría crecer muy rápidamente si no se modifica ...

... nuestras economías a crecer y desarrollarse de forma saludable ...

... que nos permitirá crecer rápidamente cuando el contexto ...

Scoring Translations: Context

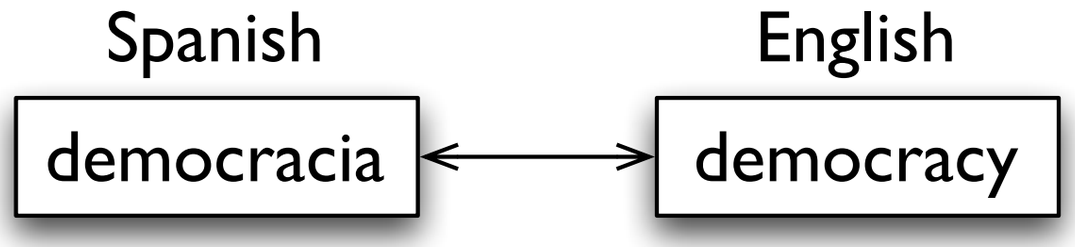


Scoring Translations: Context

eólica	estambul	admirable	choque
wind	**istanbul**	remarkable	**shock**
nuclear	virginia	wonderful	shocks
hydroelectric	zagreb	**admirable**	**clash**
geothermal	london	splendid	disagreement
photovoltaic	oreja	magnificent	disparity
purchasing	rosales	excellent	link
saving	moscow	outstanding	contradiction
efficiency	attending	fantastic	divisions
atomic	washington	producing	confrontation
wielded	johannesburg	commendable	synergies

Scoring Translations: Orthography

Etymologically related words often retain similar spelling across languages with the same writing system



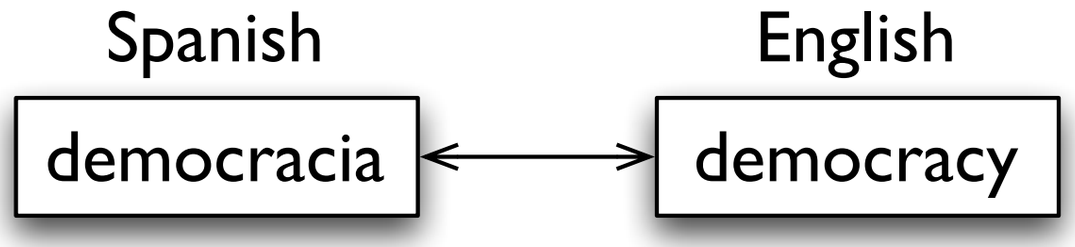
Words with lower edit distances are sometimes good translations of each other

Scoring Translations: Spelling

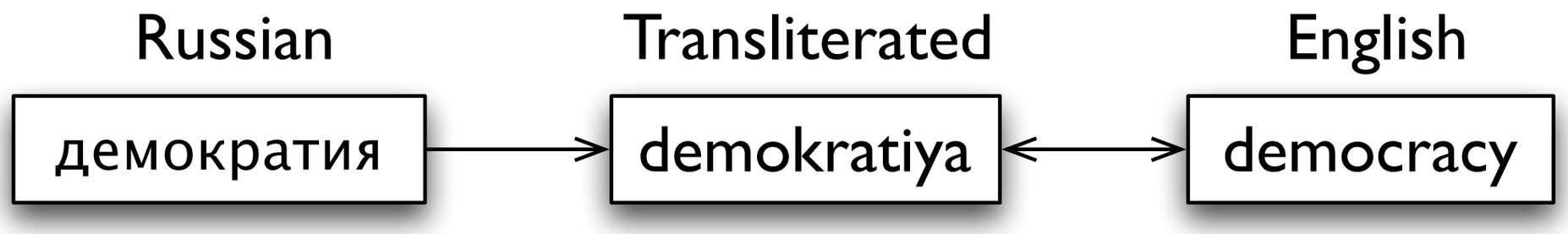
sanitario	desarrollos	volcánica	montana
sanitary	ferroalloy	**volcanic**	**montana**
sanitation	barrosos	volcanism	fontana
unitario	destroyers	voltaic	montane
sanitarium	mccarroll	vacancy	mentana
sanitation	disallows	konica	montagna
sagittario	disallow	dominica	montanha
sanitarias	scrolls	veronica	montan
kantaro	payrolls	monica	montano
sanitorium	carroll	volcano	montani
santoro	steamrolls	vratnica	montand

Scoring Translations: Orthography

Measuring edit distance for languages which share the same writing system



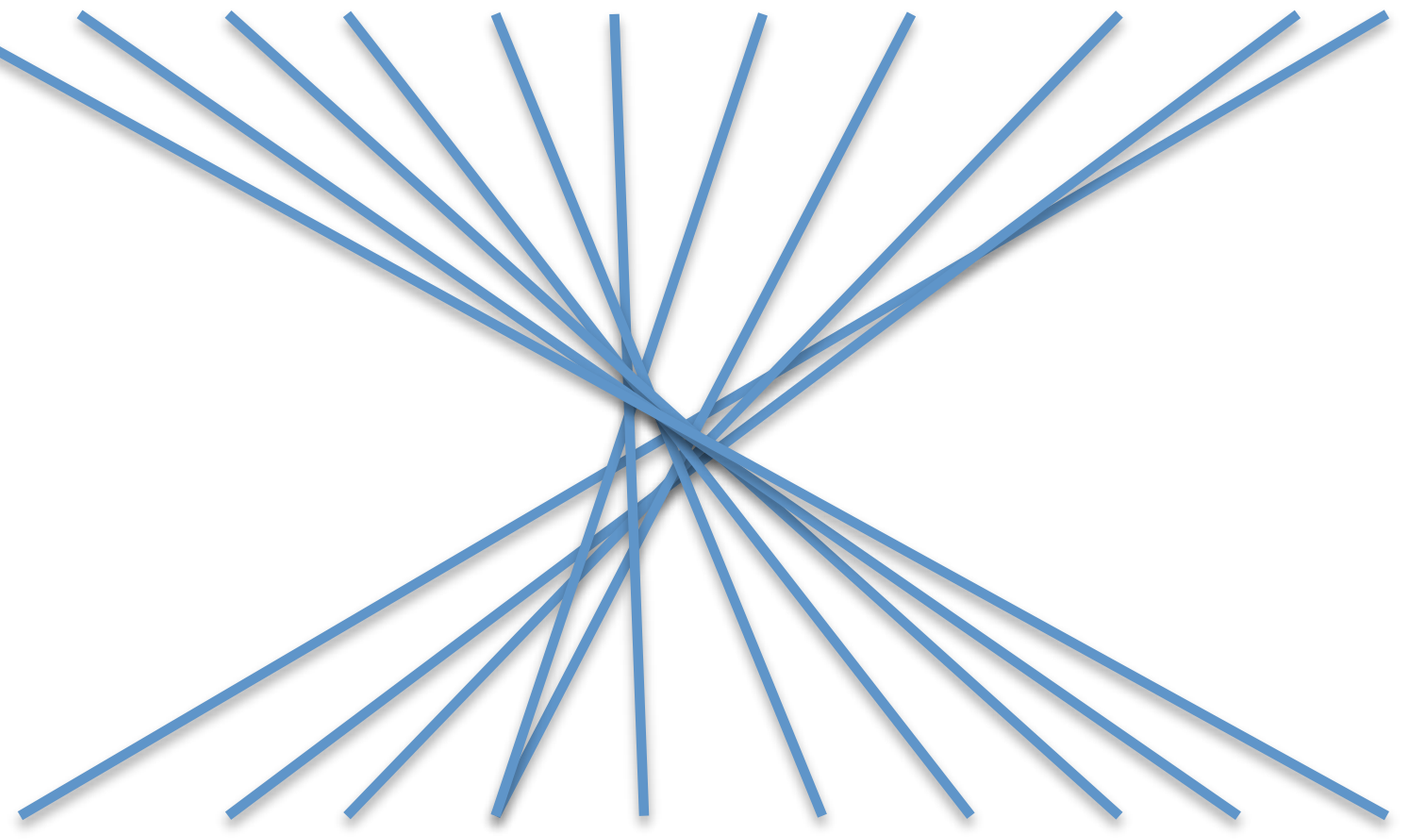
We transliterate for languages with different writing systems



Assign a similarity score with edit distance or with a discriminative transliteration model

Transliteration using SMT

ا ی ر ی ڈ ن ی ز ی گ ی ل ا



a l e x a n d r i a

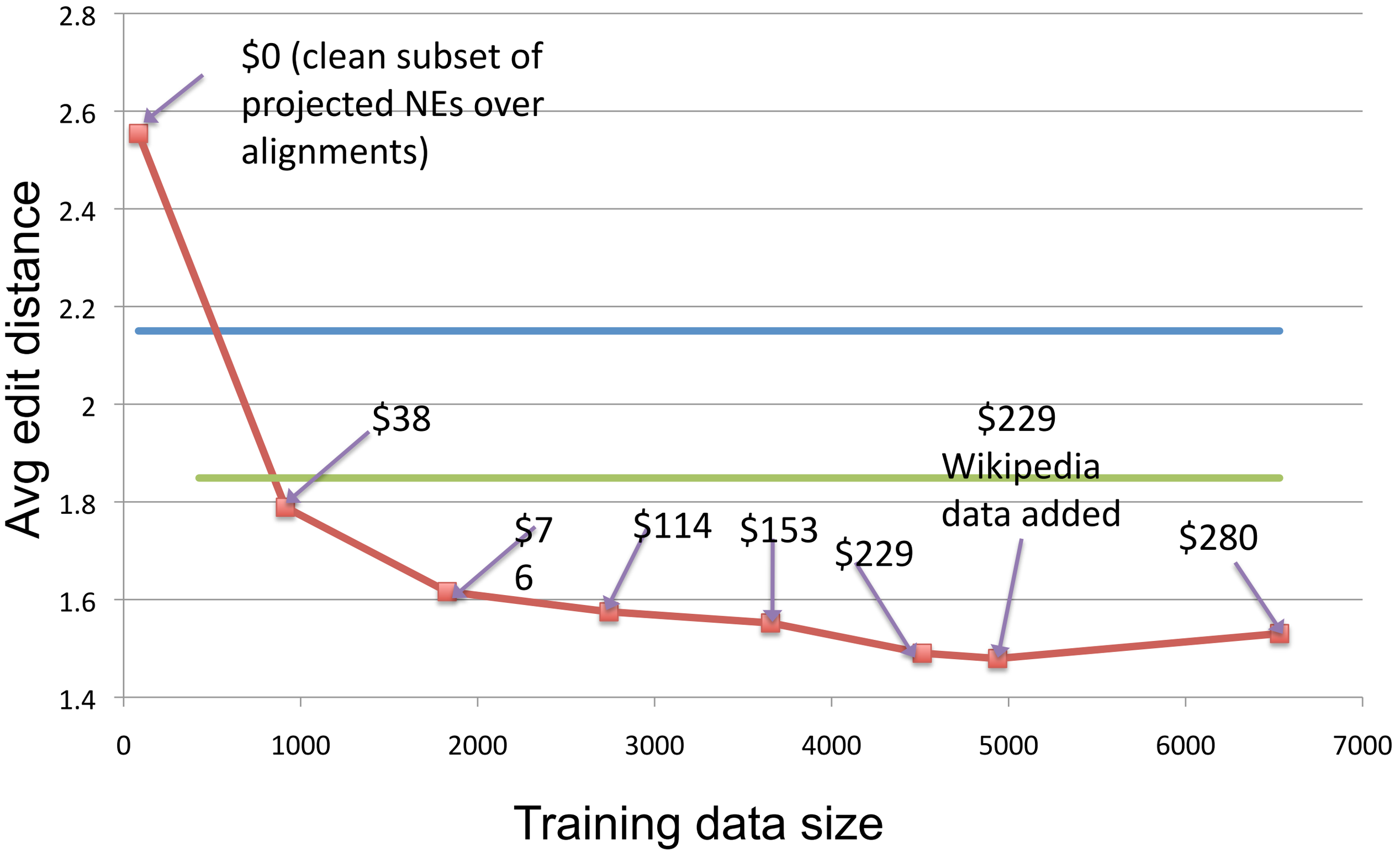
Character-based translation

- Instead of aligning words across sentence pairs, we align characters across name pairs
- Learn translation rules for sequences of letters
- Language model is n-graph letter sequence built from English names
- Requires:
  - Many pairs of foreign-English names
  - Many names written in English for LM

Transliteration training data

- Extracted name pairs from automatically word aligned parallel corpus
- Gathered training data from Wikipedia
  - 890 articles about people w/inter-language links
- Hired Urdu speakers on Mechanical Turk to transliterate names
  - gathered 5,470 English->Urdu names and 5,470 Urdu->English names
  - 2/3 of the data was high quality
  - 12,384 additional names for <$300

Learning Curve

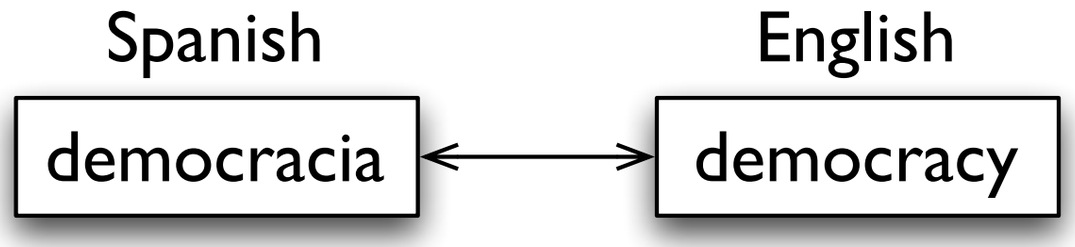


Example transliterations

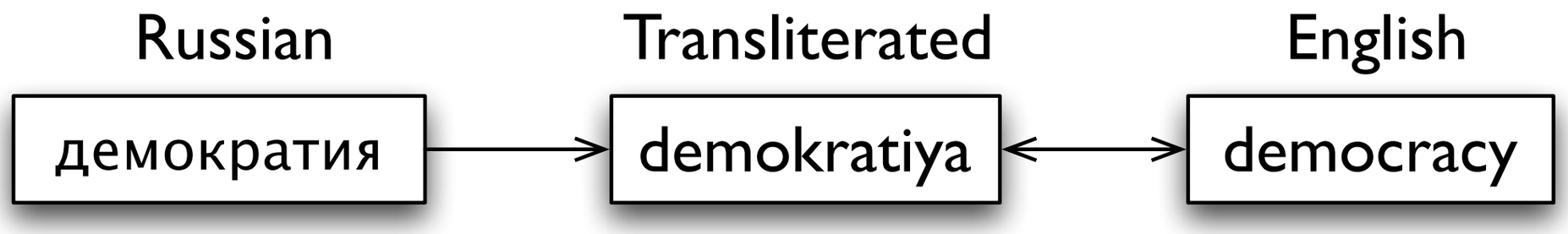
Training Data Size	Intermediate Output			
84	korشed	orosco	hu	moqtaders
914	khurshid	urska	yao	muqtadera
1828	khurshid	wersk	yao	muqtra
2742	khurshid	urska	yao	muqtara
3655	khursheed	versk	yao	muqtadera
4512	korshed	orsik	yaho	muqtadera
4938	khurshid	versk	yaho	muqtadra
6531	khursheed	warsak	yahu	moqtadar
Correct Transliteration	**khurshid**	**warsac**	**yahoo**	**muqtadra**

Scoring Translations: Orthography

Etymologically related words often retain similar spelling across languages with the same writing system



We transliterate for languages with different writing systems



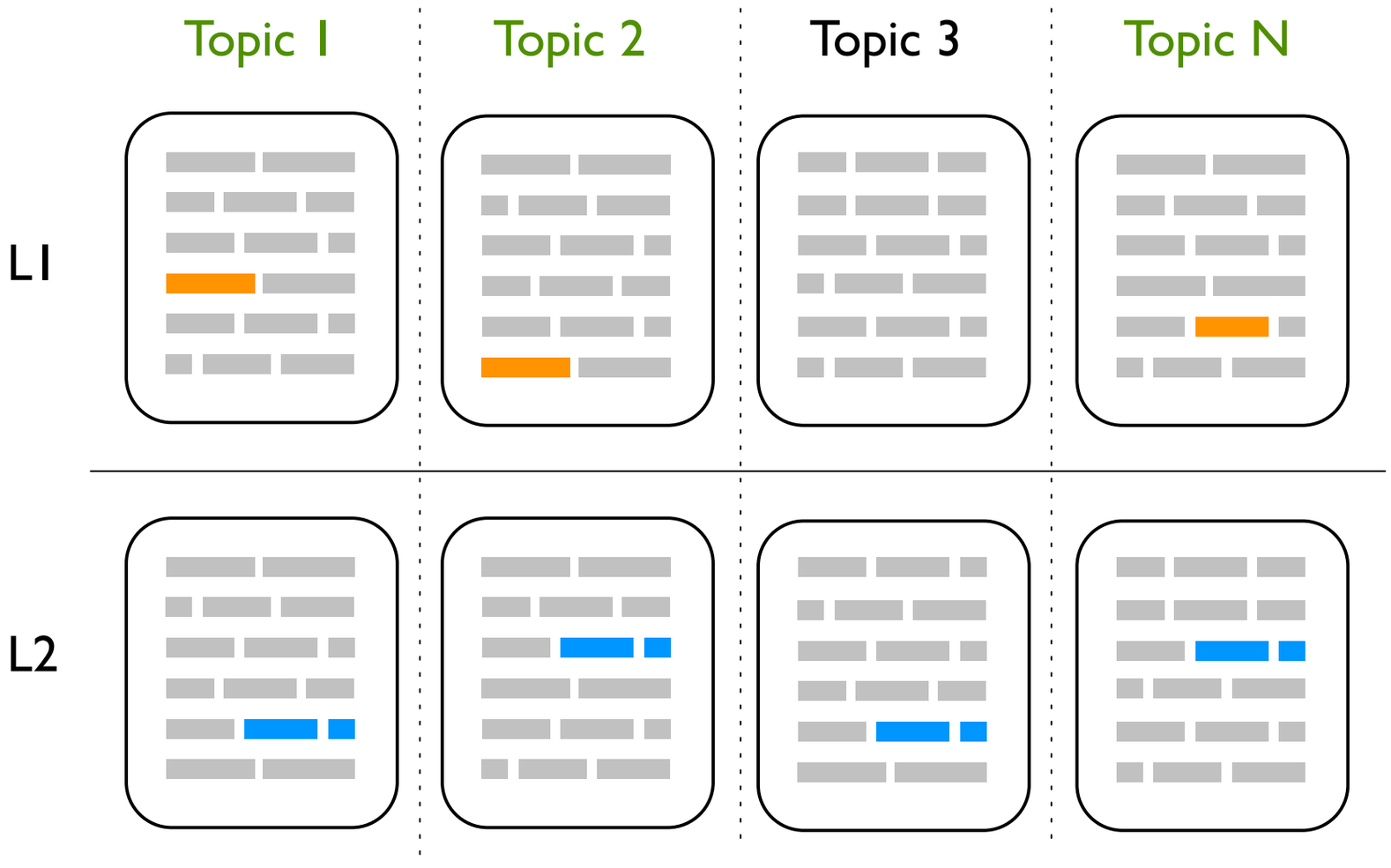
Assign a similarity score with edit distance or with a discriminative transliteration model

Scoring Translations: Topics

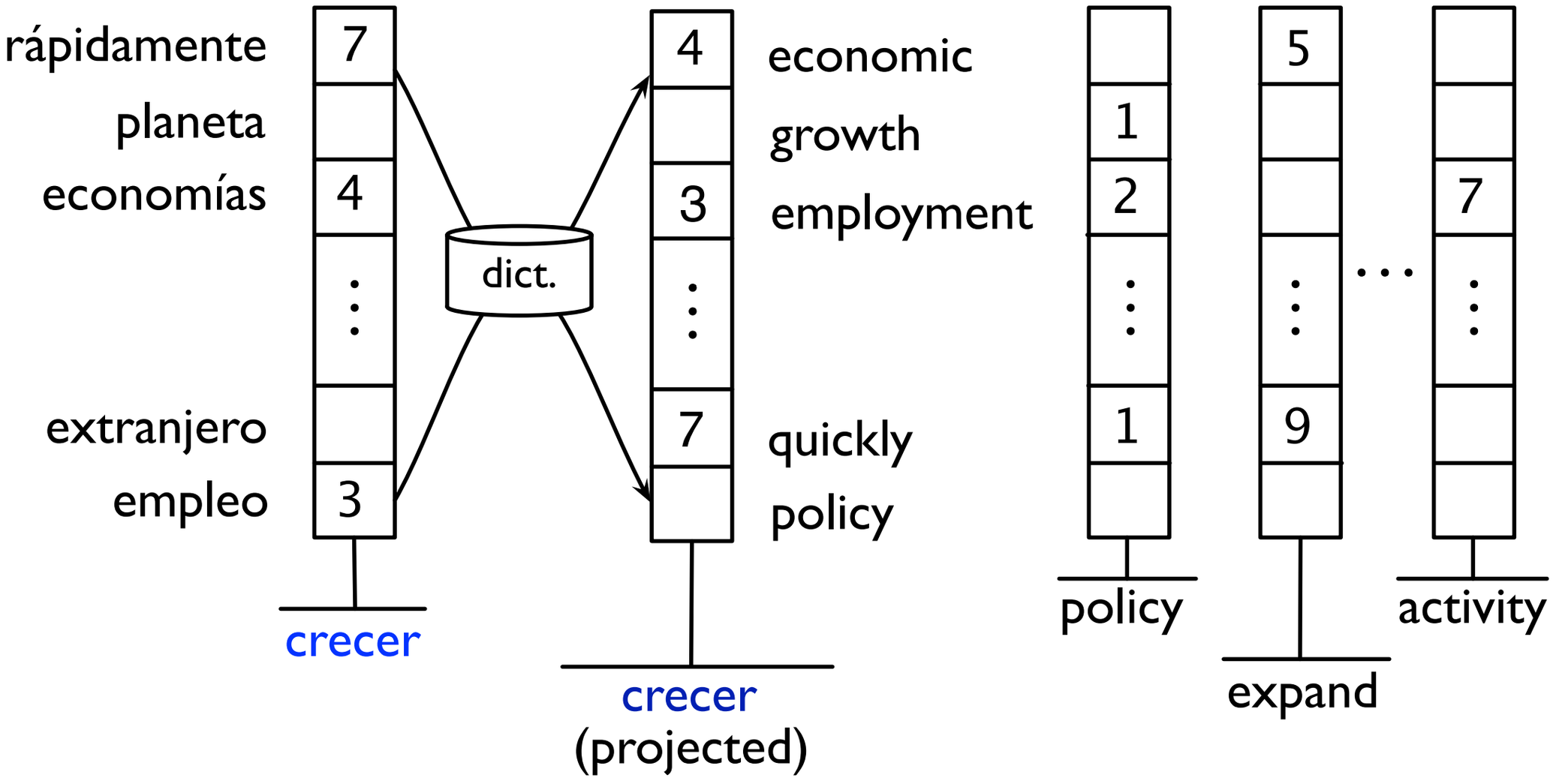
Phrases and their translations used to describe the same topics.

The more similar the set of topics two phrases appear in, the more likely they are translations.

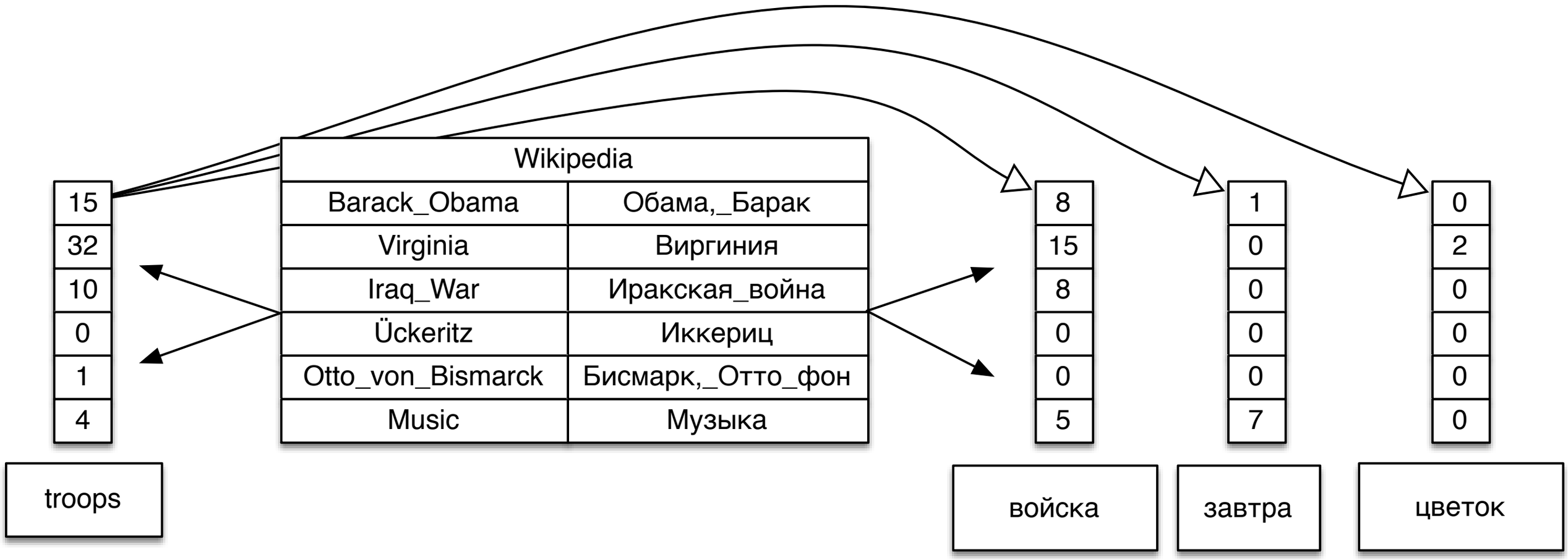
We treat Wikipedia article pairs with interlingual links as topics.



Scoring Translations: Context



Scoring Translations: Topics



Scoring Translations: Context

sanitario	**desarrollos**	**volcánica**	**montana**
health	**developments**	**volcanic**	**montana**
transcultural	developed	eruptions	miley
medical	development	volcanism	hannah
sanitation	used	lava	beartooth
patient	using	plumes	cyrus
deliverables	modern	eruption	crazier
pharmaceutica	based	volcano	bozeman
sewerage	important	volcanoes	chelsom
healthcare	history	breakouts	absaroka
care	different	volcanically	baucus

How good is each approach?

We have a wide variety of using monolingual texts to measure translation equivalence. Which is the best?

We measured the accuracy on 24 languages: *Albanian, Azeri, Bengali, Bosnian, Bulgarian, Cebuano, Gujarati, Hindi, Hungarian, Indonesian, Latvian, Nepali, Romanian, Serbian, Slovak, Somali, Swedish, Tamil, Telugu, Turkish, Ukrainian, Uzbek, Vietnamese* and *Welsh.*

For each foreign word we computed a ranked list of English words using each signal of translation equivalence.

The number of candidate English words varied by language, from 34,000 to 287,000.

How good is each approach?

We compared the predictions against a bilingual dictionary for each language, and calculated whether a good translation occurred anywhere in its its top-*k* predictions.

$$acc_k = \frac{\sum_{l \in L} I_{lk}}{|L|}$$

Sum over all test words

I_{lk}: 1 iff a correct item is in the top-*k* list of translations for word *l*

Accuracy at rank *k*

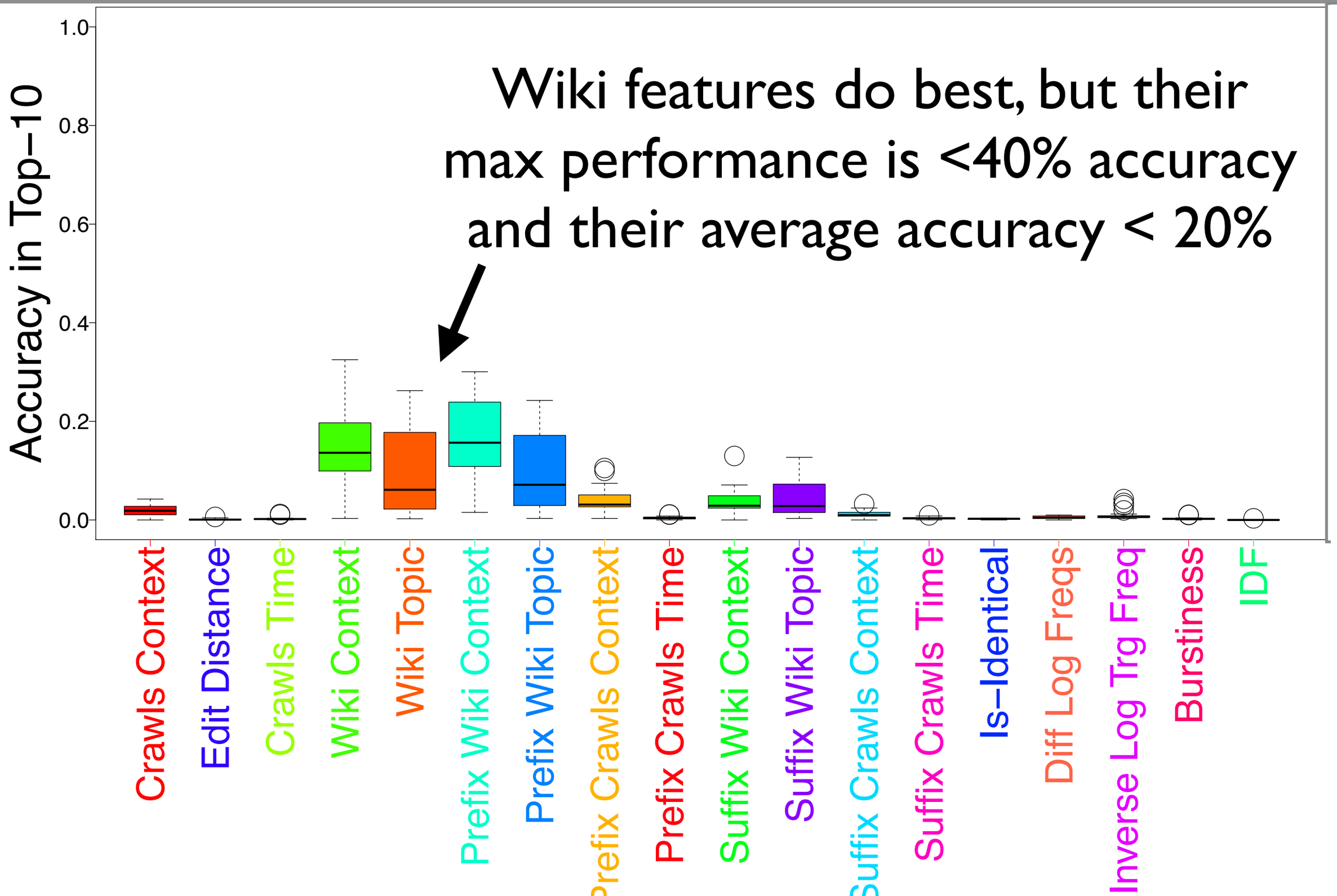
$|L|$: Number of words in the test set for a language

Wide range of signals

We measured the top-10 accuracy for 18 signals of translation equivalence, and averaged across the 24 languages.

1. Web Crawls Contextual Similarity
2. Web Crawls Temporal Similarity
3. Orthographic Similarity
4. Wikipedia Contextual Similarity
5. Wikipedia Topic Similarity
6. Wikipedia Frequency Similarity
7. Wikipedia IDF Similarity
8. Wikipedia Burstiness Similarity
9. Web Crawls Prefix Contextual Similarity
10. Web Crawls Prefix Temporal Similarity
11. Web Crawls Suffix Contextual Similarity
12. Web Crawls Suffix Temporal Similarity
13. Wikipedia Prefix Contextual Similarity
14. Wikipedia Prefix Topical Similarity
15. Wikipedia Suffix Contextual Similarity
16. Wikipedia Suffix Topical Similarity
17. String Identity
18. Inverse Log of Target Wikipedia Frequency

Per-Signal Results



Combining signals

On its own, each of these measures of translation equivalence is a weak signal.

Can we combine the weak signals into something stronger? If so, how?

$$MRR_e = \frac{\sum_{h \in H} \frac{1}{r_h(e)}}{|H|}$$

1 over the rank

Mean Reciprocal Rank

Set of signals

Rank of word *e* by signal *h*

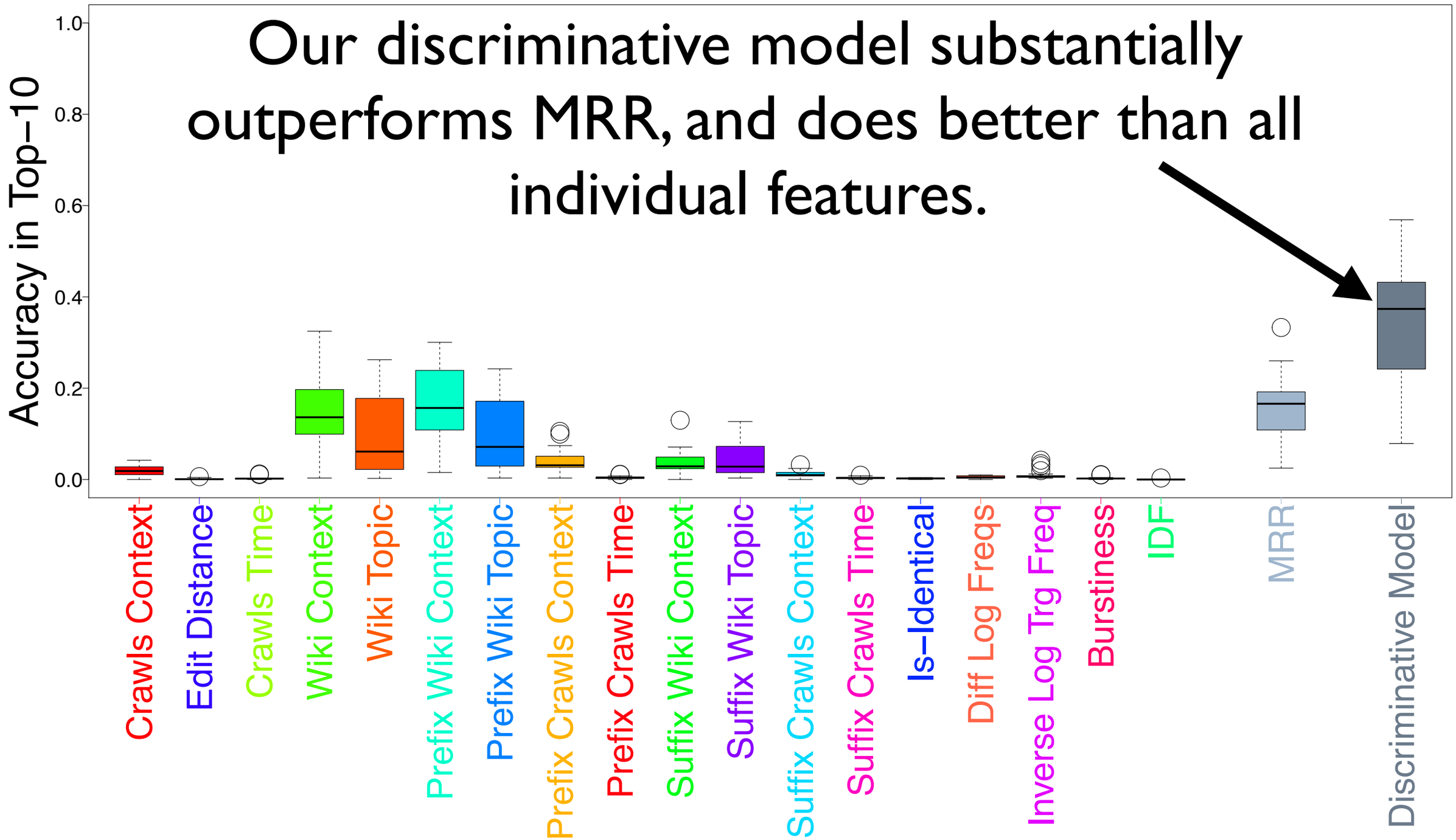
Combining signals

MRR is an unsupervised approach to combining signals. We also introduce a novel discriminative approach that exploits the fact we use a small bilingual dictionary to project across vector spaces.

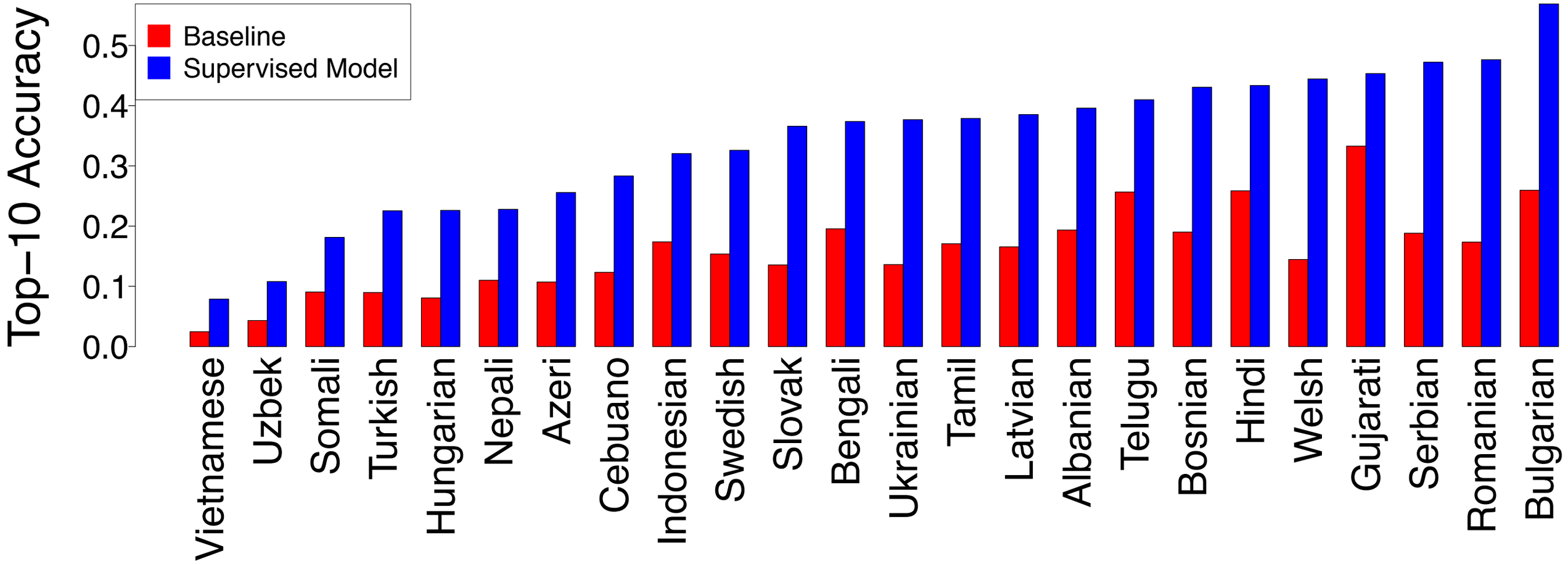
We train a binary classifier to predict whether a word is a translation or not. Translations from our dictionary serve as positive training examples. Each one is paired with 3 randomly selected non-translations as negative training examples.

We rank translations based on the strength of the classifier's prediction that a word is a translation.

Per-Signal Results



Per-Language Results



Example translation

Dictionary gloss for Hindi Wikipedia article

one forest one उ&च density its साथ one field is wood (tree) एसज5गल its lots definitions , is which of various मानद5डo on based J.यह पोदाM total ९.४ % the earth of surface को surround R is (either 30 %) those of आवासo (habitat) STUोलोiगक Vवाह (hydrologic flow) मोWलातोसX (modulator) , and soil (soil) safeguard , one the earth its बीओ]फअ का rules important sides of गठन.का foreign do is history telling is , of " forest " one बीहड़ field whose means कानaनी for on बाजa of for nidhirit ◌िशकार (hunting) its iारा साम5ती (feudal) कuलीनता (nobility) is , and these ◌िशकार in jungles compulsory more if me all (see wild no was royal forest (royal forest)) .हालsiक , ◌िशकार its in jungles usual वuडलuड its importance areas को शिामल did while , श।vद forest at the end wild land more generally means do of for was था.एक वuडलuड (woodland) which of one ज5गल from different is .

Example translation

Translation for same article with Dictionary + Transliterations + Induced Translations

one forest one systolic density of which one field is tree (tree) canopy of many definitions , is which of various crm on based han.yh nearly headless . % of the earth surface ko surround te is (or 30 %) which of keyhole (organisms) canopy irr (telecom low) modulators (coniferous) , and soil (erosion) safeguard , one the earth of app ka more important sides of gthn.ka foreign to do is history telling is , of " forest " one maestra field whose means responsibility for on pulleys of for nidhirit mane (africana) of dhara necker (electors) émigrés (forest) is , and these lions forests more necessary if among all (see no wild the royal forest (royal society)) .hallanki , mane of forests often evergreen of important areas ko they did while , quirk forest at the end wild land more generally means do its for was tho.aq evergreen (forests) which of one forest from different is .

End-to-End MT

Could we do full end-to-end machine translation without using any bilingual parallel corpora?

Aside from learning the translations of words, and estimating their probabilities, what else would we need?

Discuss with your neighbor.

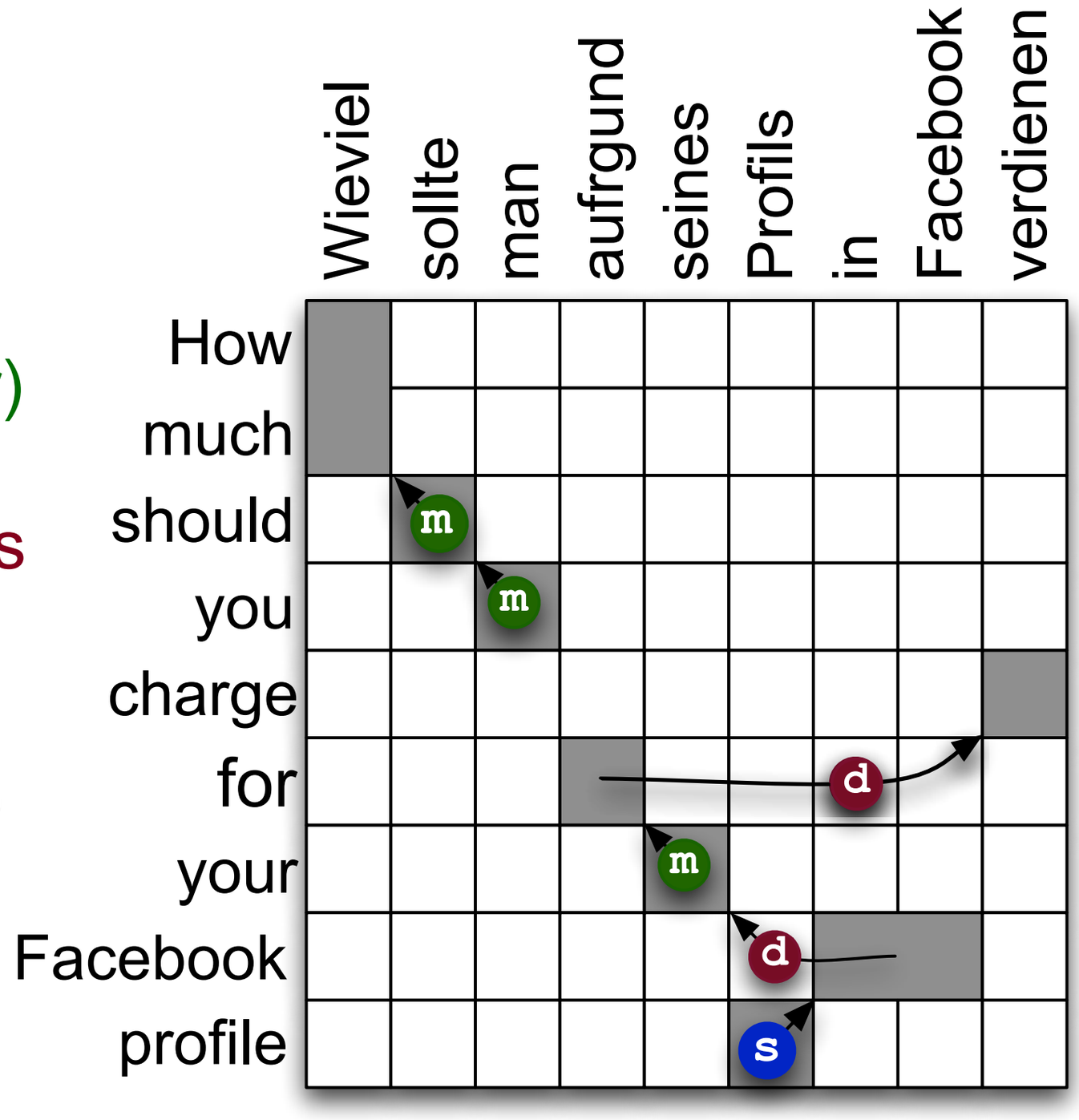
Re-ordering model

m: monotone (keep order)

s: swap order

d: become discontinuous

Reordering features are probability estimates of s, d, and m

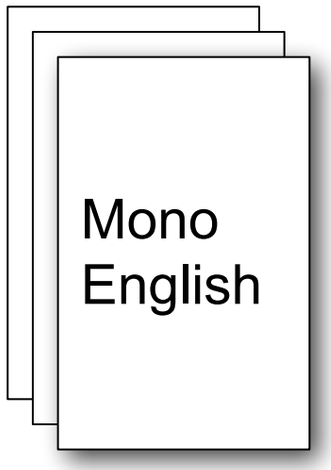


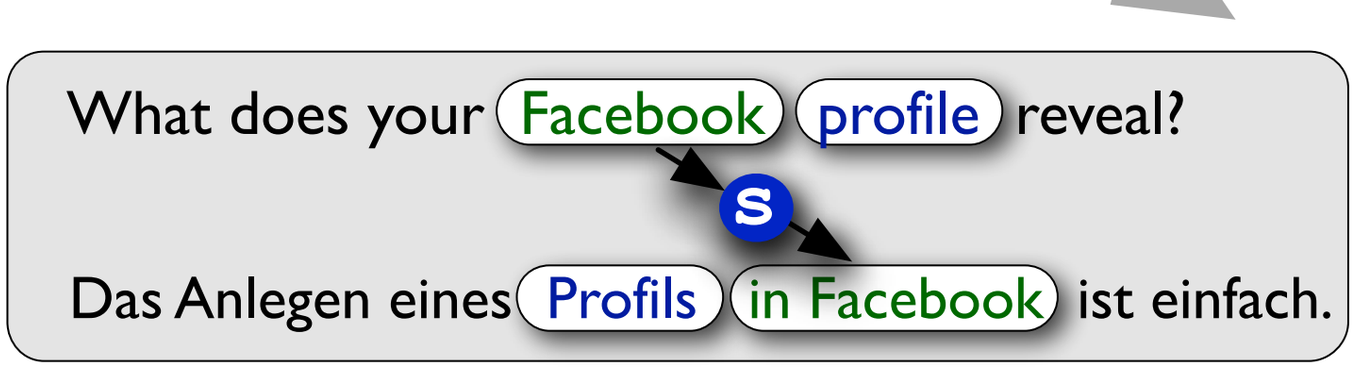
Re-ordering model (monolingual)

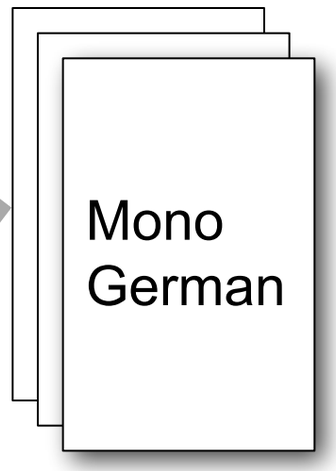
Estimate same probabilities, but from pairs of (unaligned) sentences taken from monolingual data

Phrase Table

German	English
! das	, and
profile	Profils
…	…
Facebook	in Facebook
…	…
und nicht	and a lack
zustand	situation as

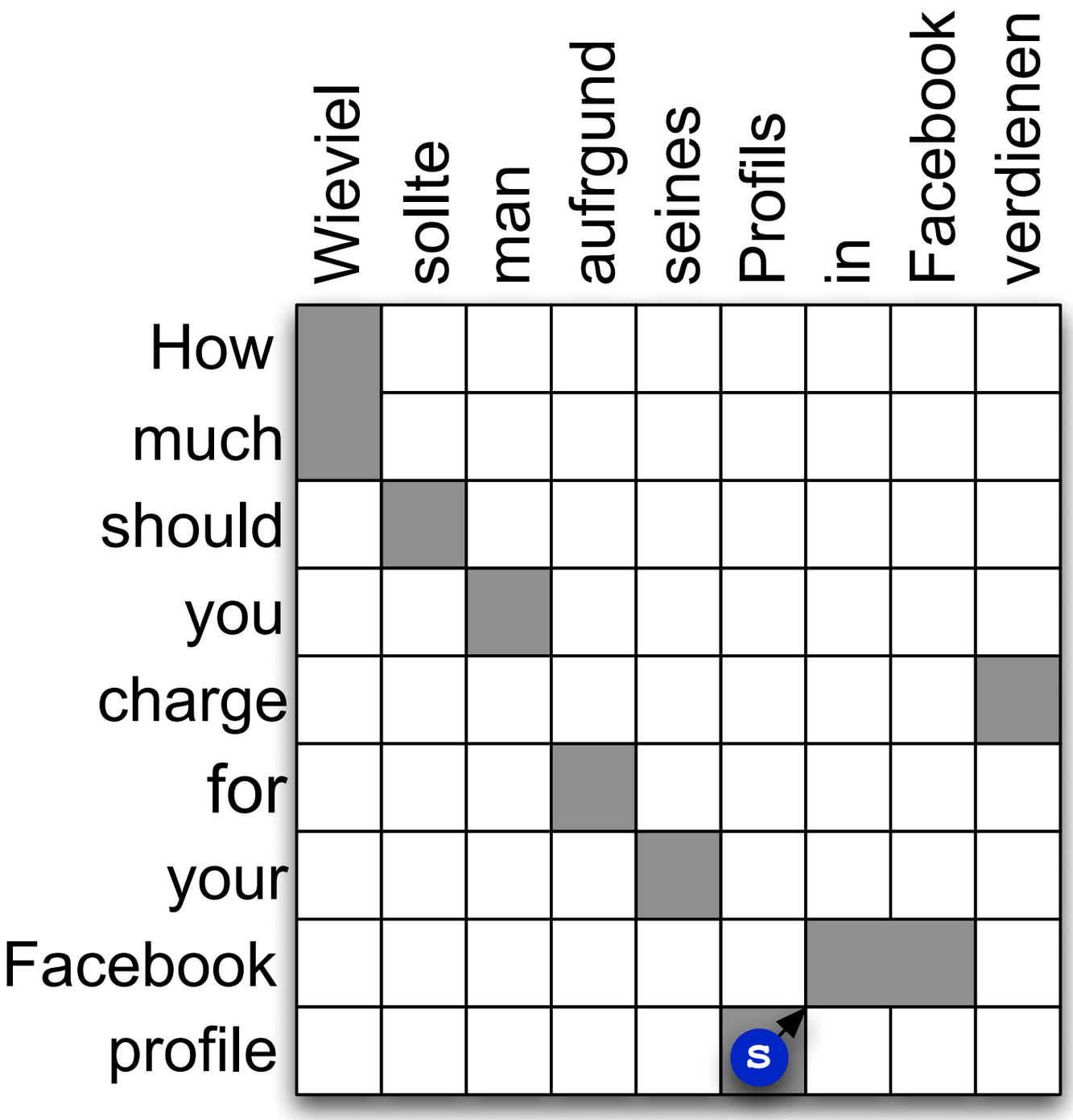


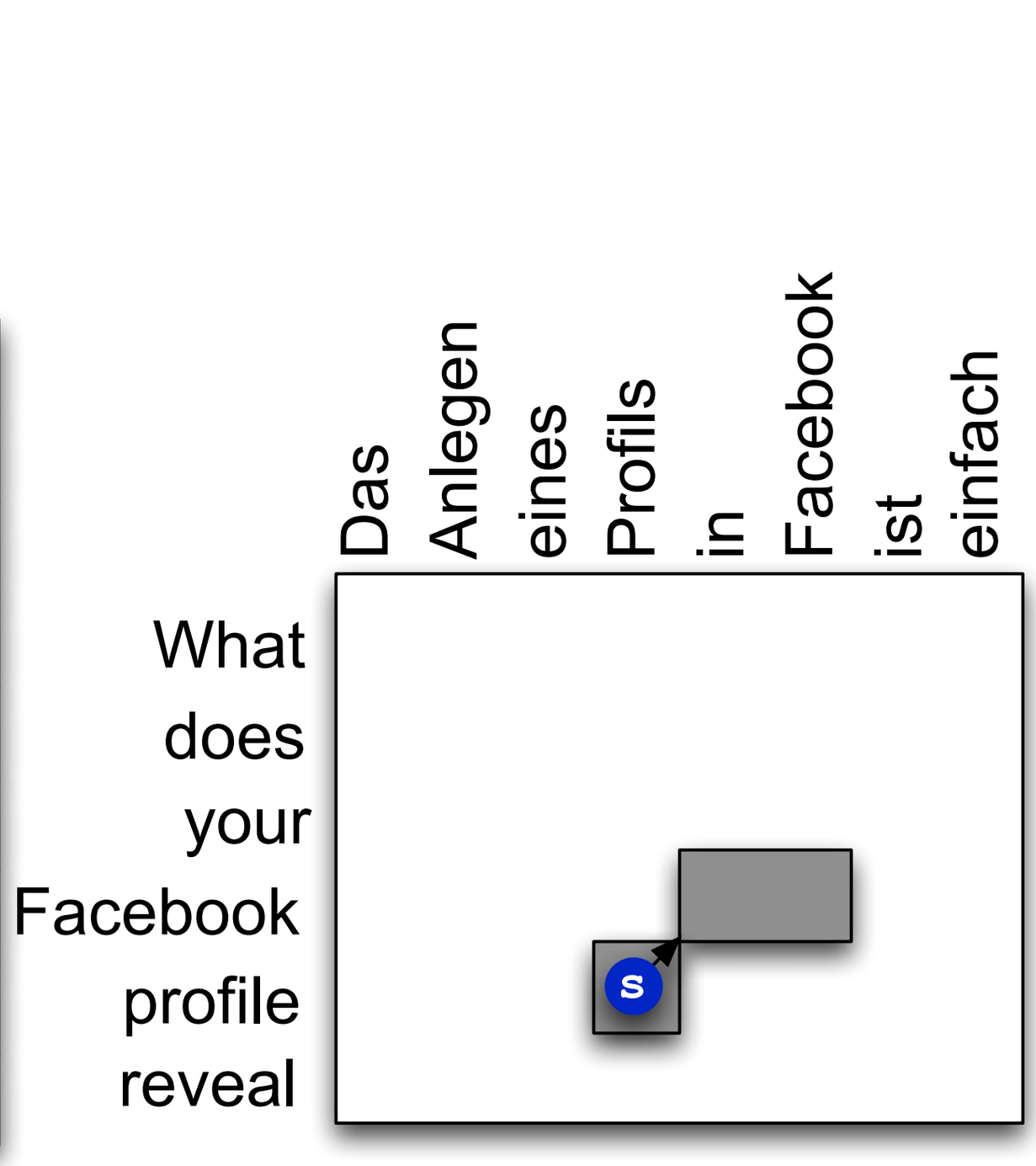




Repeat over many sentences

Re-ordering model



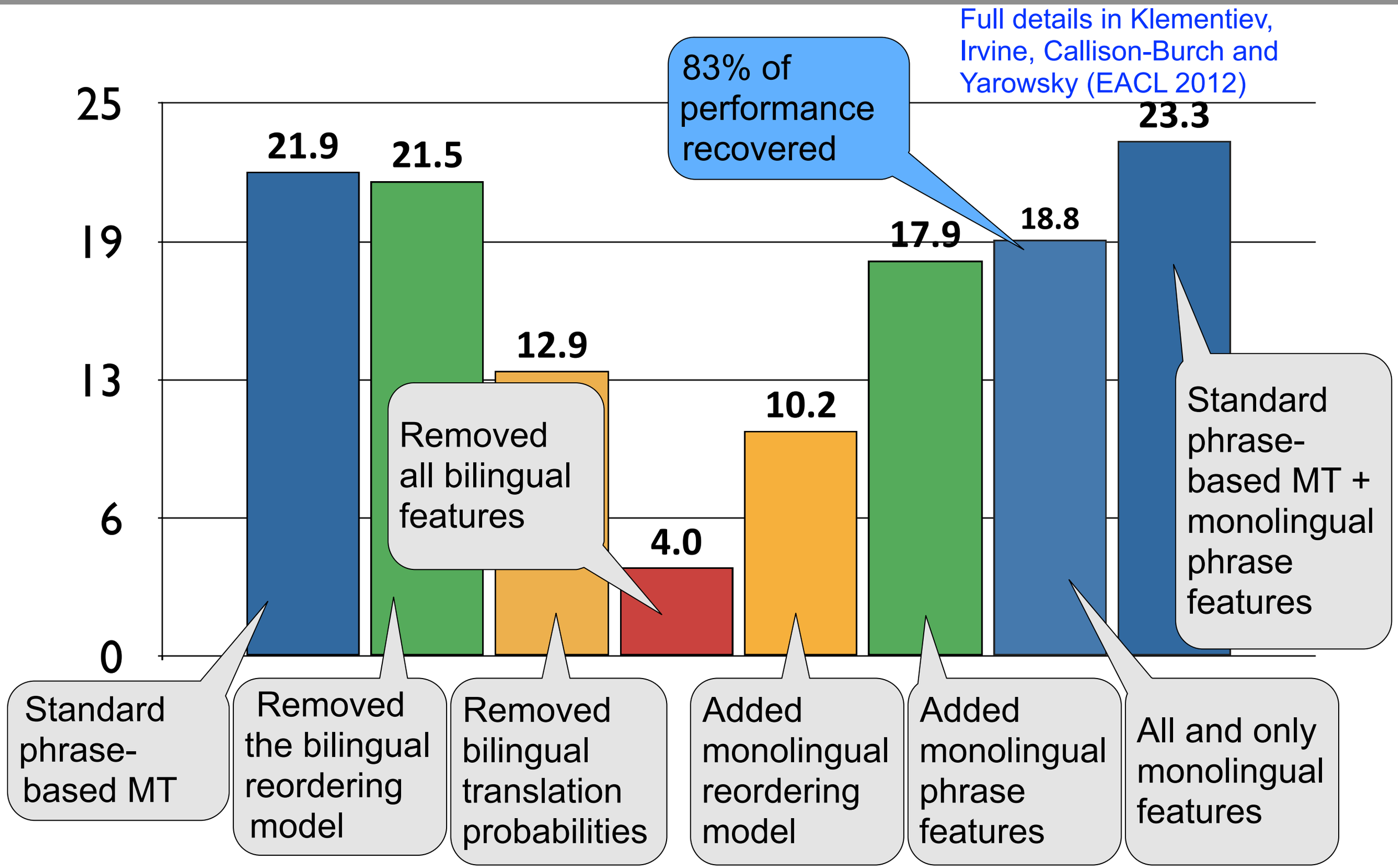


Experimental Setup

- How well can we estimate the parameters a phrase-based SMT system with monolingual data?
  - Performed ablation study to removed each part of the standard bilingually estimated system
  - Restored each component with monolingual equivalent
- Cleanroom experiment
  - Phrase-table is same across two conditions
- Data
  - Europarl parallel corpus (50M words)
  - Spanish and English Gigaword corpora (1B words)
  - Spanish and English paired Wikipedia articles (40-60M words)

Idealization

Spanish-English MT w/o bitexts



Translation comparison

Bilingually estimated

The US administration can inject 700 billion dollars in banking

The highest representatives of the congress and the government, the president George W. Bush, reached agreement in a pact in broad terms of financial aid to the system of American finance. The vote will take place at the beginning of next week.
The American legislators caused a gap in the talks on the approval of the rescue plan in the form of aid to the US financial system with the amount of 700 billion dollars.
However, is not yet won.
The US congressmen must fine-tune certain details of the contract before they can make public the final shape of the law and that is adopted.

Monolingually estimated

The US government can inject 700 billion dollars of the bank

The highest representatives of congress and the government, the president George W. Bush, agreed to a pact many terms of financial aid to the system of finance American. The vote will take place as early as next week.
The legislature American caused a breach in talks on the approval of rescue plan in the form of the financial system American with the amount of 700 million dollars.
However, is not yet livestock .
Congress further to some details of the contract before it can make public the final form of the law, with an voted.

Announcements

Will Lewis from Microsoft Research will be giving the lecture on Thursday. He has a lot of job openings. If you'd like to meet with him, email me tonight.

Deadlines:
Tonight - complete term project is due. No extensions

April 16 - read over other students' projects and vote on the ones you want to do as your final HW assignment.

Tuesday April 28th (last day of class):
(1) Turn in your solution to one of the other team's projects as your final HW.
(2) Language research assignment is due