Toward Statistical Machine Translation without Parallel Corpora

Alex Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky
Johns Hopkins University

EACL 2012

Motivation

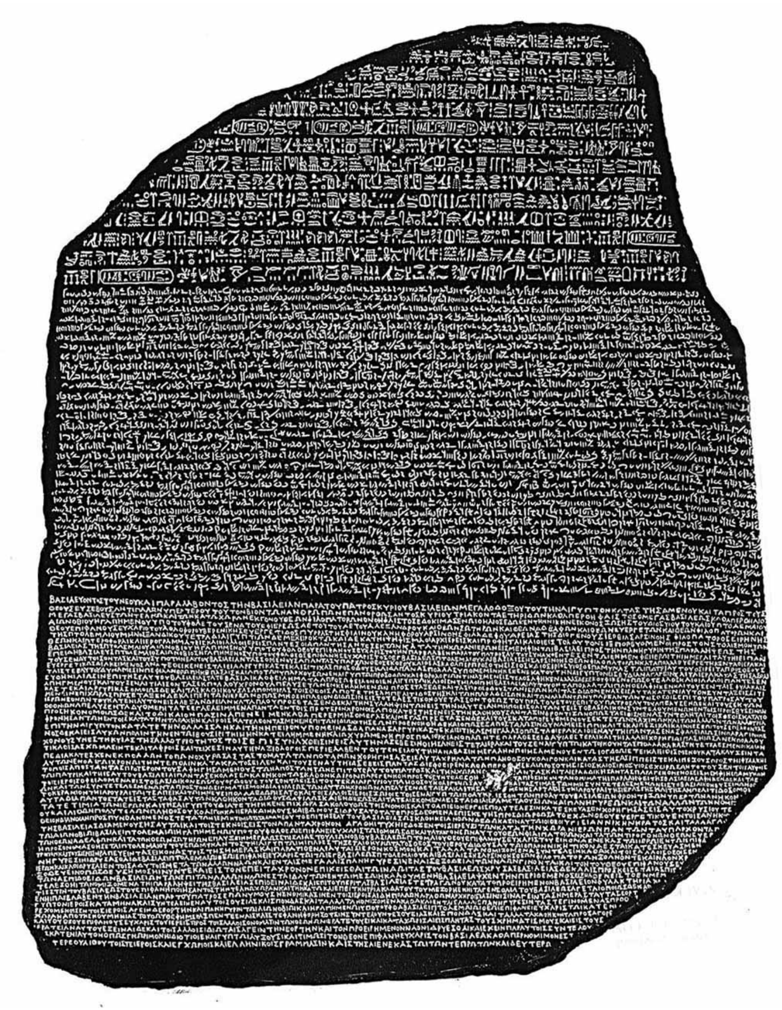
- State-of-the-art statistical machine translation models are estimated from parallel corpora (manually translated text)
- Large volumes of parallel text are typically needed to induce good models

BUT

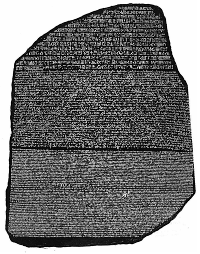
- Manual translation is laborious and expensive
- Sufficient quantities available for only a few language pairs
  - E.g. Canadian and European parliamentary proceedings
  - WMT 2011 translation task: 6 languages, Google Translate - 59

Goal

Instead of:



Goal



+

Cheap monolingual data

Use cheap and plentiful monolingual data to reduce (eliminate?) the need for parallel data to induce good translation models

Summary of the Approach

At a very high level, translation involves:

1. Choosing correct translation of words and phrases

Phrase based SMT: extract / score phrasal dictionaries from sentence aligned parallel data

This work: extend bilingual lexicon induction to score phrasal dictionaries from monolingual data

2. Putting the translated phrases or words in the right order

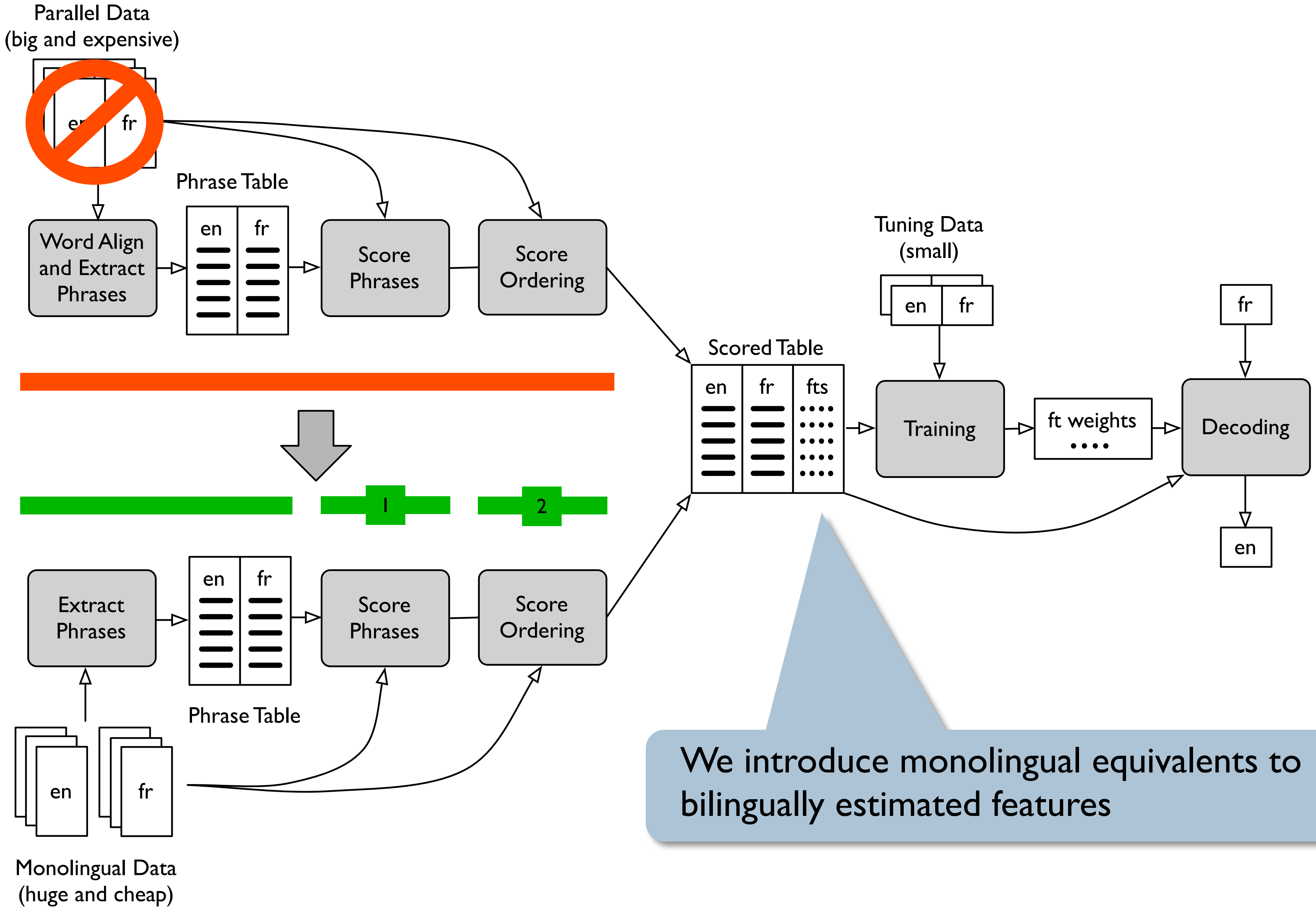
Phrase based SMT: extract ordering information from parallel data

This work: novel algorithm to estimate ordering from monolingual data

Outline

- Motivation and summary of the approach
- Phrase-based statistical machine translation
- Weaning phrase-based SMT off parallel data
  - Scoring phrases
    - Extending bilingual lexicon induction
  - Scoring ordering
    - Novel reordering algorithm
- Experiments

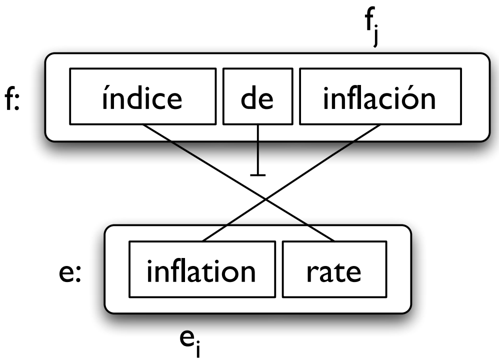
Phrase-based SMT



We introduce monolingual equivalents to bilingually estimated features

Scoring Phrases

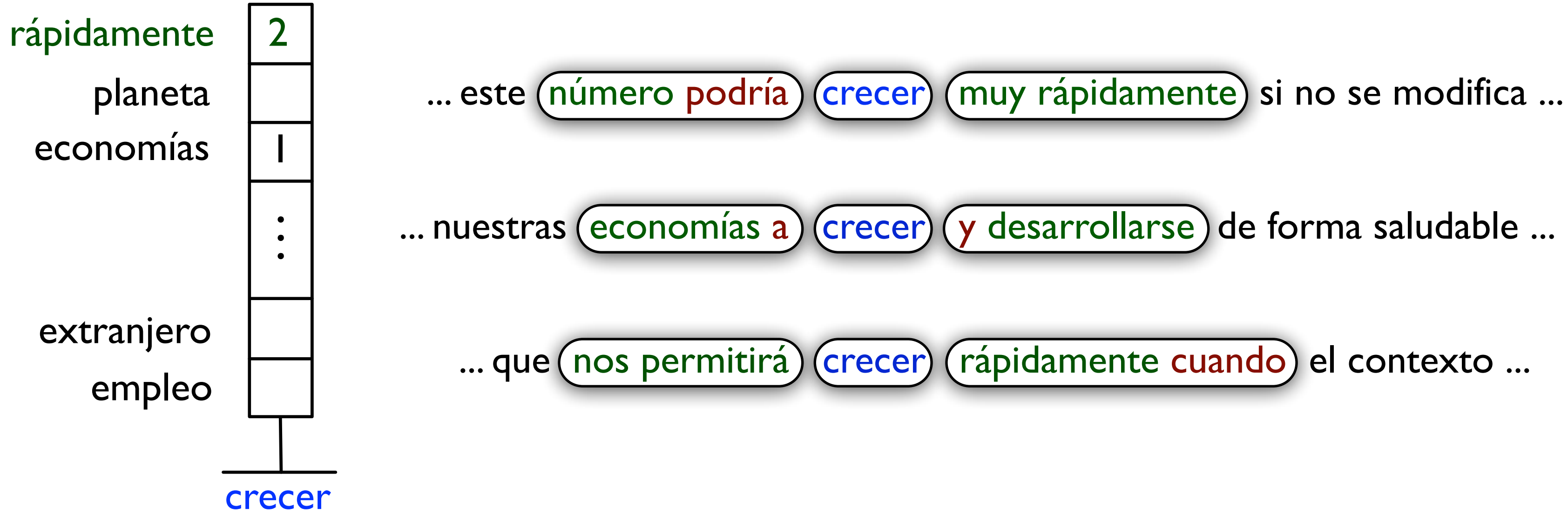
- Phrase-based SMT: relatedness of a given phrase pair *(e,f)* is captured by:
  - Phrase translation probabilities $\phi(e|f)$, $\phi(f|e)$
  - Average word translation $w(e_i|f_j)$ probabilities using phrase-pair-internal word alignments



  - Both estimated from a parallel corpus
- How can we induce phrasal similarity from monolingual data?

I Scoring Phrases: Context

- An old first idea [Rapp, 99]: measure contextual similarity
  - Words appearing in similar context are probably related
  - First, collect context



rápidamente 2
planeta
economías 1
⋮
extranjero
empleo
crecer

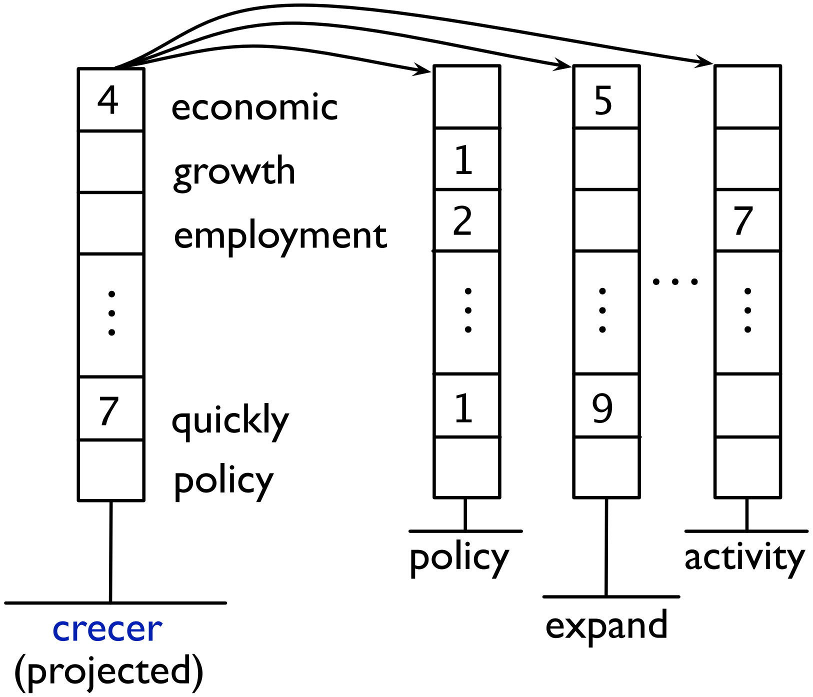
... este número podría crecer muy rápidamente si no se modifica ...

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

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

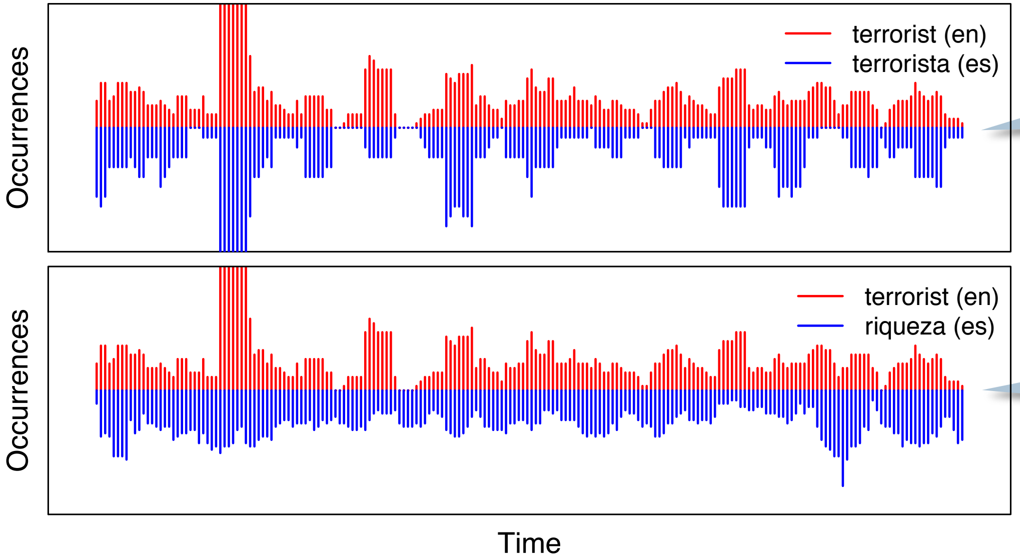
Scoring Phrases: Context

- An old first idea [Rapp, 99]: measure contextual similarity
  - Words appearing in similar context are probably related
  - First, collect context
  - Then, project through a seed dictionary, and compare vectors



I Scoring Phrases: Time

- Second idea: measure temporal similarity
  - Assume, we have temporal information associated with text (e.g. news publication dates)
  - Events are discussed in different languages at the same time
  - Collect temporal signature



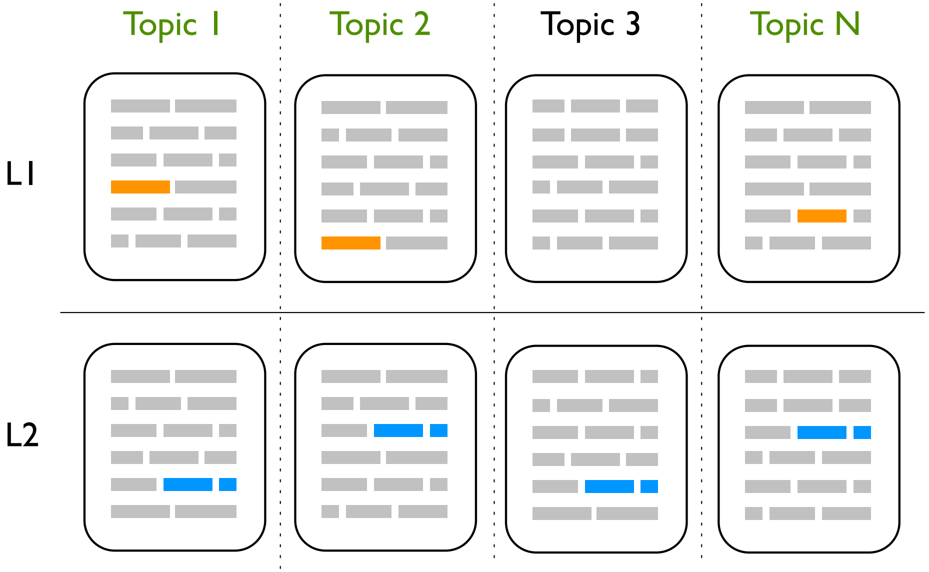
Similar

Dissimilar

- Measure similarity between signatures (e.g. cosine or DFT-based metric)

Scoring Phrases: Topics

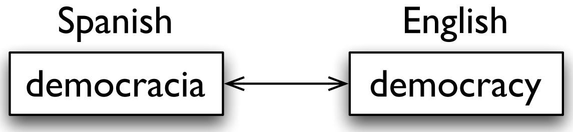
- Third idea: measure topic similarity
  - Phrases and their translations are likely to appear in the same topic
  - The more similar the set of topics in which a pair of phrases appears the more likely the phrases are similar



- We treat Wikipedia article pairs with interlingual links as topics

Scoring Phrases: Orthography

- Fourth idea: measure orthographic similarity
  - Etymologically related words often retain similar spelling across languages with the same writing system



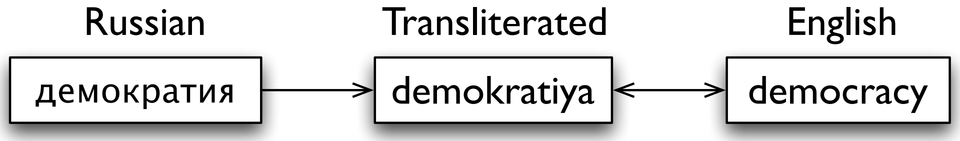
Spanish

democracia

English

democracy

  - Transliterate for language pairs with different writing system



Russian

демократия

Transliterated

demokratiya

English

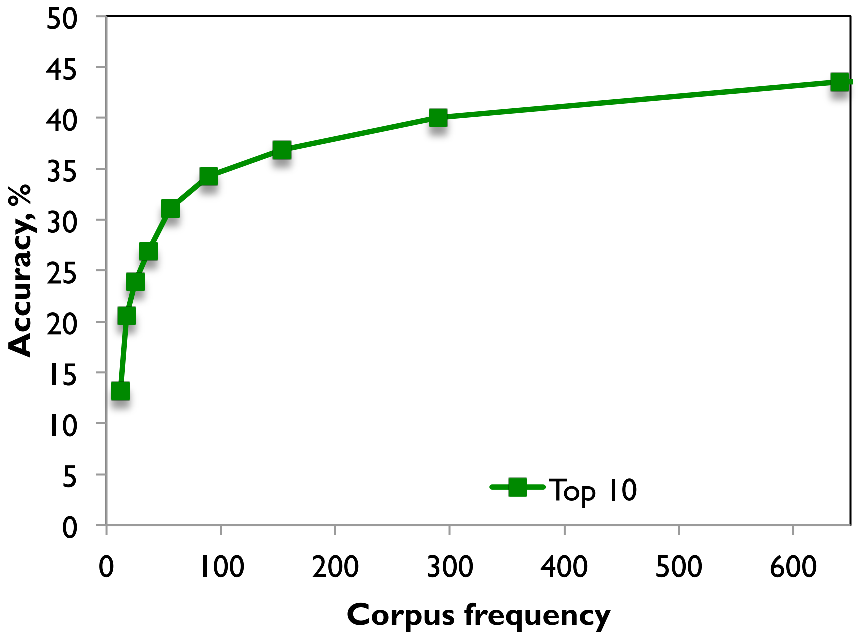
democracy

  - Measure similarity with edit distance or a discriminative translit model

- Other ideas: burstiness, etc.

I Scoring Phrases: Scaling Up Lex Induction

- Challenge in scoring phrase pairs monolingually: Sparsity

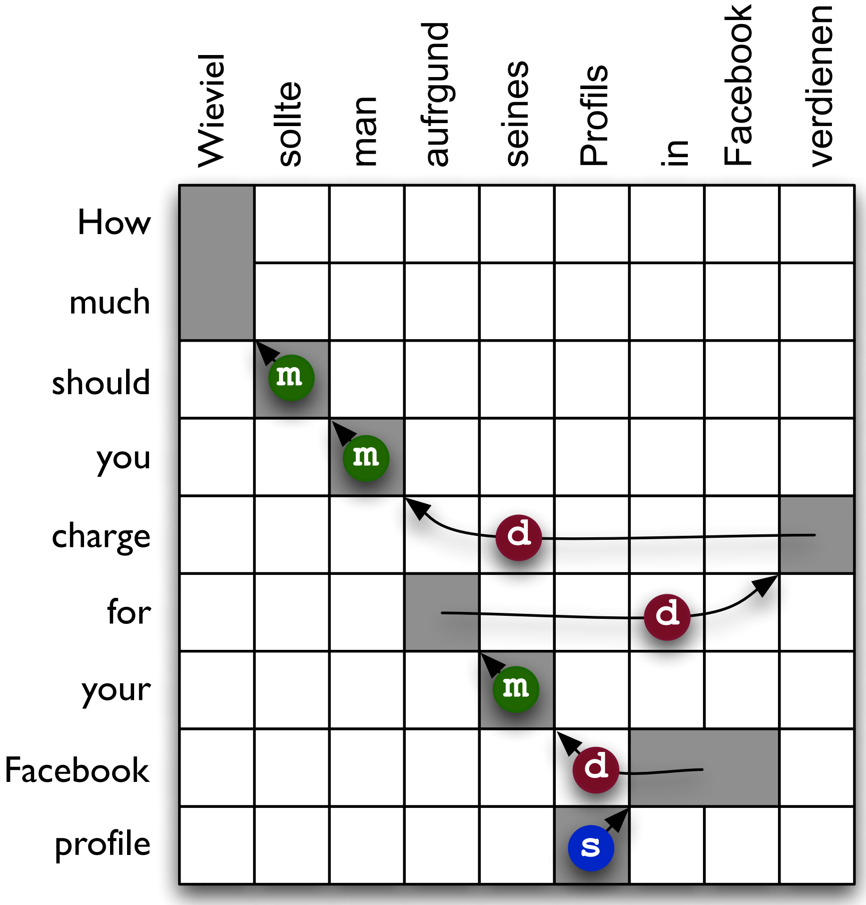


- Score the similarity of individual words within a phrase pair
  - Similar to lexical weights in phrase-based SMT
  - Use phrase pair internal word alignments, penalize for unaligned words

2 Scoring Ordering

- Phrase-based SMT: model the probabilities of orientation change of a phrase with respect to a preceding phrase when translated

    - Start with aligned phrases
    - **m**: monotone (keep order)
    - **s**: swap order
    - **d**: become discontinuous
    - Reordering features are probability estimates of **s**, **d**, and **m**

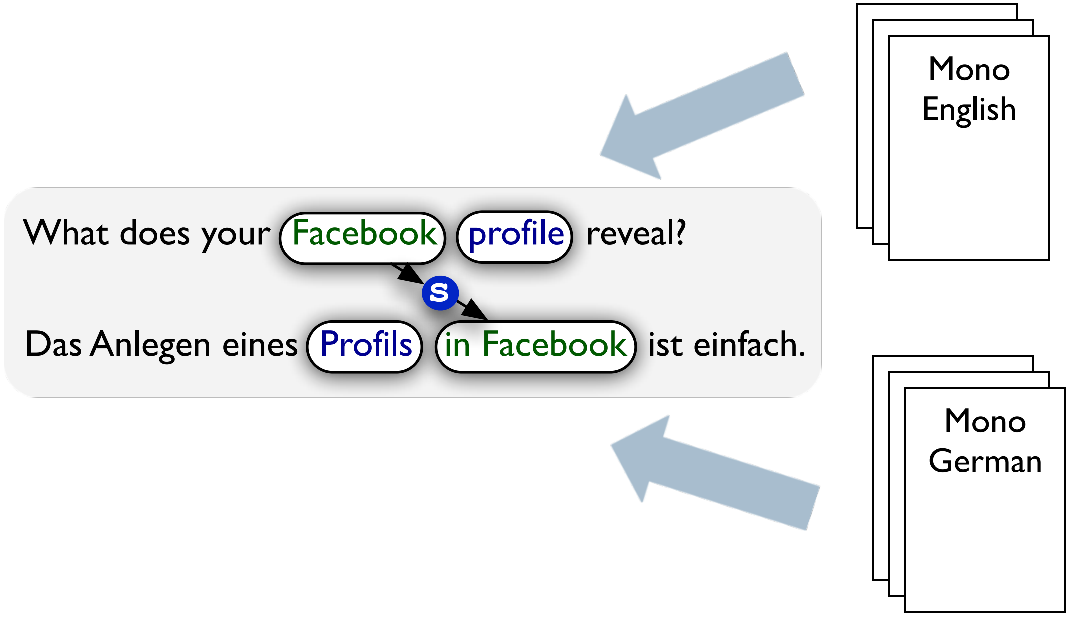


2 Scoring Ordering from Monolingual Data

- Estimate same probabilities, but from pairs of (unaligned) sentences taken from monolingual data
  - We don't have alignments, but we do have a phrase table

Phrase Table

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



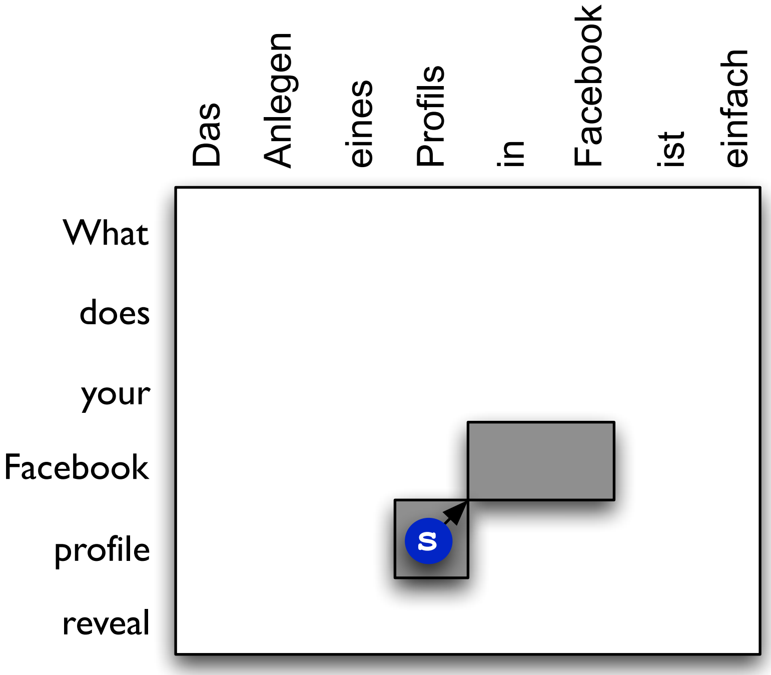
What does your Facebook profile reveal?

Das Anlegen eines Profils in Facebook ist einfach.

  - Repeat over many sentences

2 Scoring Ordering from Monolingual Data

- Estimate same probabilities, but from pairs of (unaligned) sentences taken from monolingual data
  - We don't have alignments, but we do have a phrase table



  - Repeat over many sentences

1,2 Scoring Phrases and Ordering: Summary

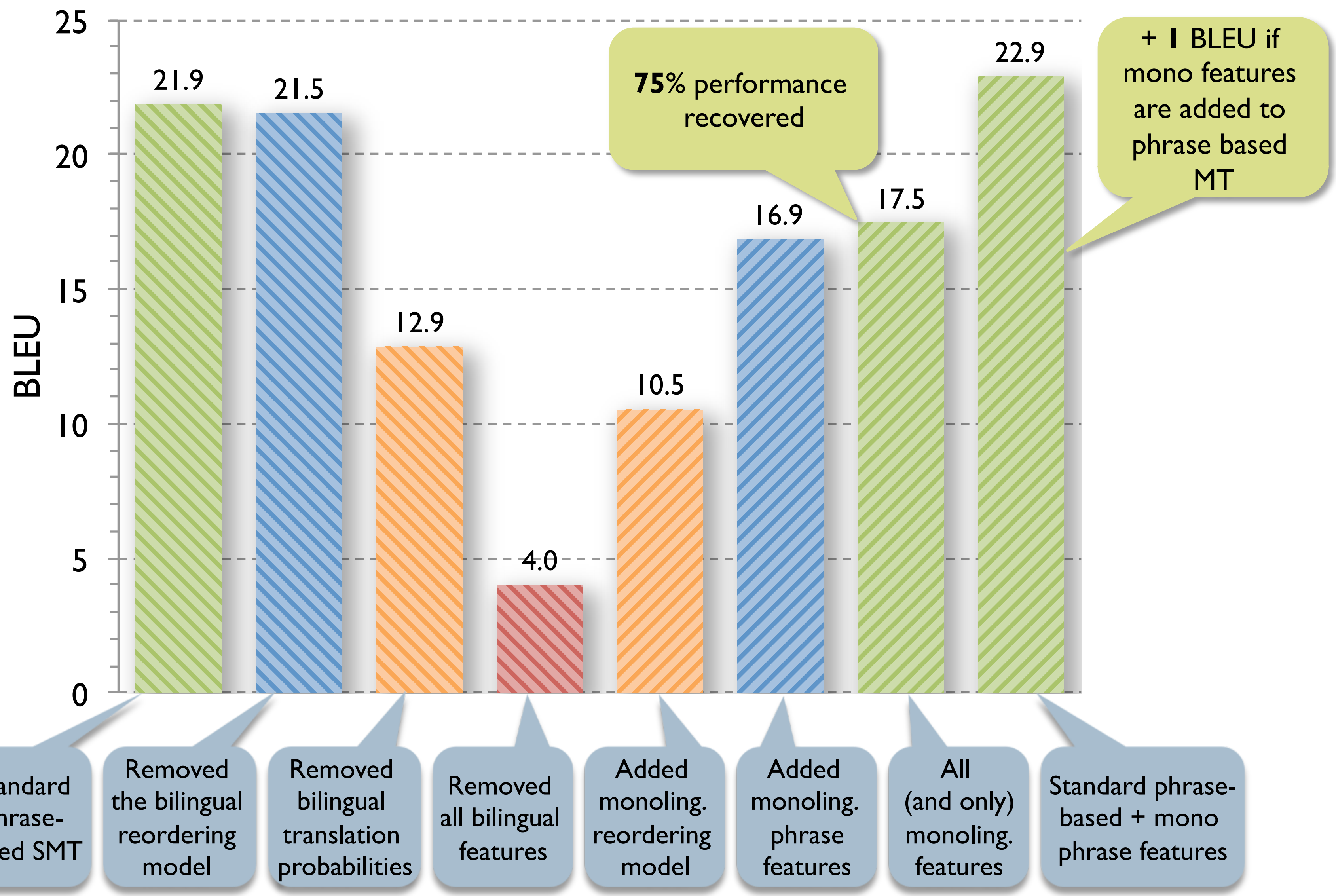
Phrase-based SMT Features	Monolingual Equivalents
Phrase translation probabilities	Temporal, contextual, and topic phrase similarity
Lexical weights	Temporal, contextual, topic, and orthographic word similarities
Reordering features	Reordering features estimated from monolingual data

Alternatives to features estimated from parallel corpora: we use cheap monolingual data (along with some metadata), and a seed dictionary

Experiments: Spanish-English

- Idealized setup: treat English and Spanish sections of Europarl as two independent monolingual corpora
  - Europarl is annotated with temporal information
- Drop the idealization: estimate features from truly monolingual corpora
  - Gigaword, Wikipedia for contextual and temporal
  - Wikipedia for topical
- Run a series of lesion experiments:
  - Begin with standard features estimated from parallel data
  - Drop phrasal, lexical, and reordering features
  - Replace them with the monolingually estimated counterparts
  - See how much performance can be recovered

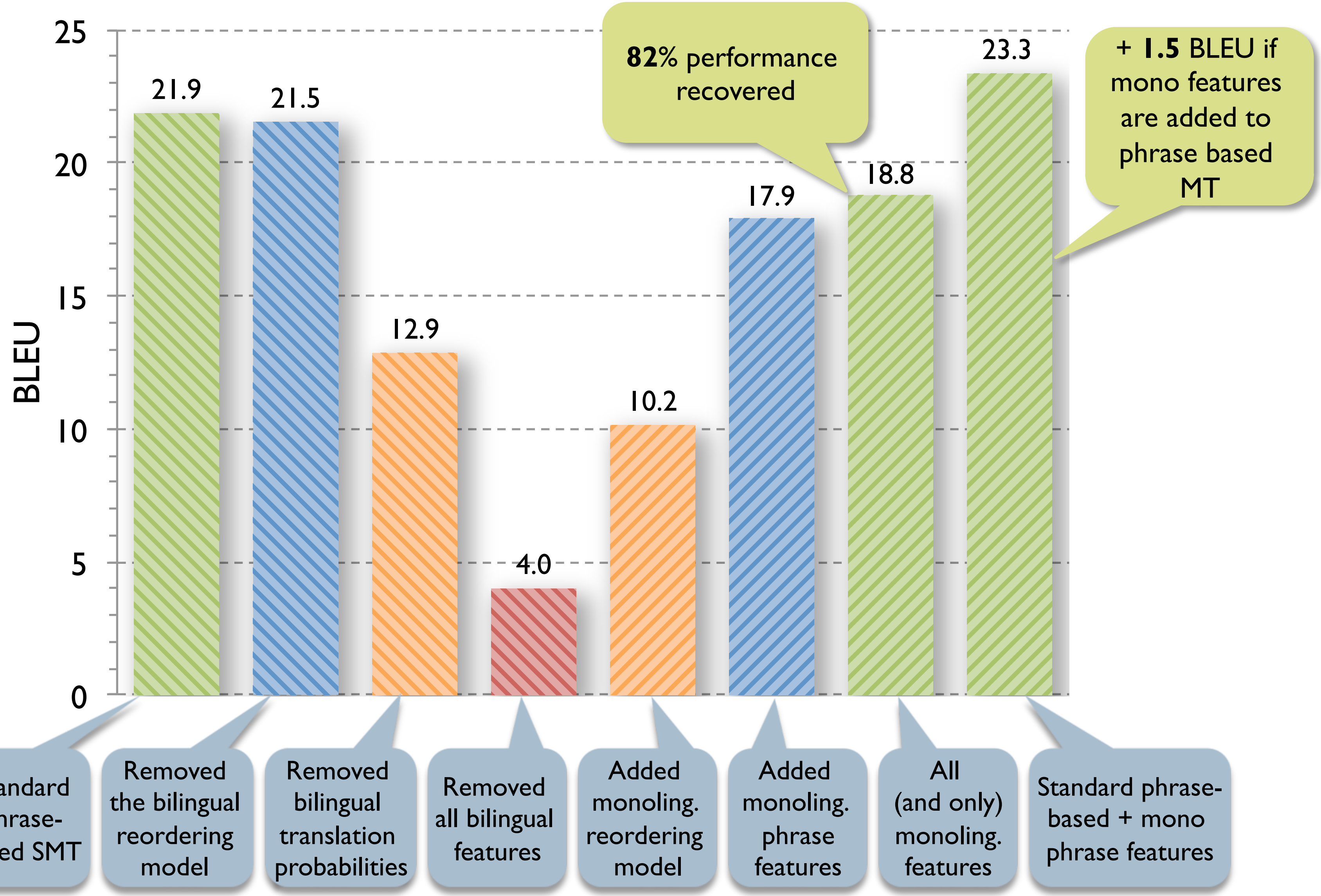
Experiment: Europarl Spanish-English



75% performance recovered

+ **1** BLEU if mono features are added to phrase based MT

Experiment: Monolingual Spanish-English



82% performance recovered

+ **1.5** BLEU if mono features are added to phrase based MT

Standard phrase-based SMT

Removed the bilingual reordering model

Removed bilingual translation probabilities

Removed all bilingual features

Added monoling. reordering model

Added monoling. phrase features

All (and only) monoling. features

Standard phrase-based + mono phrase features

Conclusions

Monolingual data takes us a long way toward inducing good translation models

- Can recover substantial portion of the lost performance in lesion experiments
- Consistently improve performance when added to the phrase-based setup

Important direction for languages lacking sufficient (or any) parallel data

- True for most languages
- Monolingual data is plentiful and growing