Machine Translation

Natural Language Processing
CS 4120/6120—Fall 2016
Northeastern University

David Smith
some slides from
Charles Schafer & Philip Koehn
Translation and NLP

- Translation is one of the oldest language tasks tried on a computer
  - Just look at that archaic name: “Machine Translation”!
- Translation involves many linguistic systems
- “Apollo program” dual-use argument:
  - Translation models of alignment and transfer are useful in question answering, paraphrase, information retrieval, etc.
interlingua

source text

analysis

transfer

generation

direct translation

target text
Overview

• What problems does MT address? What does it (currently) not address?
• Models: What makes a good translation?
• Alignment: Learning dictionaries from parallel text
• Next: non-parallel text, translation decoding and training
The Translation Problem and Translation Data
The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশেষ শান্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি
The Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশেষ শাস্তি, স্বাধীনতা এবং ন্যায়বিচারের ভিত্তি।
The Translation Problem

Whereas recognition of the inherent dignity and of the
equal and inalienable rights of all members of the
human family is the foundation of freedom, justice and
peace in the world
Why Machine Translation?

* Cheap, universal access to world’s online information regardless of original language. (That’s the goal)

Why Statistical (or at least Empirical) Machine Translation?

* We want to translate real-world documents. Thus, we should model real-world documents.

* A nice property: design the system once, and extend to new languages automatically by training on existing data.

\[ F(\text{training data, model}) \rightarrow \text{parameterized MT system} \]
Ideas that cut across empirical language processing problems and methods

Real-world: don’t be (too) prescriptive. Be able to process (translate/summarize/identify/paraphrase) relevant bits of human language as they are, not as they “should be”. For instance, genre is important: translating French blogs into English is different from translating French novels into English.

Model: a fully described procedure, generally having variable parameters, that performs some interesting task (for example, translation).

Training data: a set of observed data instances which can be used to find good parameters for a model via a training procedure.

Training procedure: a method that takes observed data and refines the parameters of a model, such that the model is improved according to some objective function.
Most statistical machine translation (SMT) research has focused on a few “high-resource” languages (European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world’s languages found on the web.
Most statistical machine translation research has focused on a few high-resource languages (European, Chinese, Japanese, Arabic).

Chinese  Arabic

French

Italian  Danish  Finnish

Various Western European languages: parliamentary proceedings, govt documents (~30M words)

Bible/Koran/Book of Mormon/Dianetics (~1M words)

Nothing/Univ. Decl. Of Human Rights (~1K words)

Serbian  Bengali  Uzbek  Chechen  Khmer
Most statistical machine translation (SMT) research has focused on a few “high-resource” languages (European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world’s languages found on the web.

We’ll discuss this briefly
The Translation Problem


What to translate? The most common use case is probably document translation.

Most MT work focuses on sentence translation.

What does sentence translation ignore?
  - Discourse properties/structure.
  - Inter-sentence coreference.
Sentence Translation

- SMT has generally ignored extra-sentence structure (good future work direction for the community).

- Instead, we’ve concentrated on translating individual sentences as well as possible. This is a very hard problem in itself.

- Word translation (knowing the possible English translations of a French word) is not, by itself, sufficient for building readable/useful automatic document translations – though it is an important component in end-to-end SMT systems.

Sentence translation using only a word translation dictionary is called “glossing” or “gisting”.
We’ll come back to this later...

and address learning the word translation component (dictionary) of MT systems without using parallel text.

(For languages having little parallel text, this is the best we can do right now)
Sentence Translation

- Training resource: parallel text (bitext).

- Parallel text (with English) on the order of 20M-200M words (roughly, 1M-10M sentences) is available for a number of languages.

- Parallel text is expensive to generate: human translators are expensive ($0.05-$0.25 per word). Millions of words training data needed for high quality SMT results. So we take what is available. This is often of less than optimal genre (laws, parliamentary proceedings, religious texts).
<table>
<thead>
<tr>
<th>French, English from Bitext</th>
<th>Closely Literal English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le débat est clos.</td>
<td>The debate is closed.</td>
</tr>
<tr>
<td>The debate is closed.</td>
<td></td>
</tr>
<tr>
<td>Accepteriez-vous ce principe ?</td>
<td>Accept-you that principle?</td>
</tr>
<tr>
<td>Would you accept that principle ?</td>
<td></td>
</tr>
<tr>
<td>Merci, chère collègue.</td>
<td>Thank you, dear colleague.</td>
</tr>
<tr>
<td>Thank you, Mrs Marinucci.</td>
<td></td>
</tr>
<tr>
<td>Avez-vous donc une autre proposition ?</td>
<td>Have you therefore another proposal?</td>
</tr>
<tr>
<td>Can you explain ?</td>
<td></td>
</tr>
</tbody>
</table>

(from French-English European Parliament proceedings)
Sentence Translation: examples of more and less literal translations in bitext

Word alignments illustrated. Well-defined for more literal translations.

Le débat est clos.
The debate is closed.

Accepteriez-vous ce principe?
Would you accept that principle?

Merci, chère collègue.
Thank you, Mrs Marinucci.

Avez-vous donc une autre proposition?
Can you explain?
Translation and Alignment

- As mentioned, translations are expensive to commission and generally SMT research relies on already existing translations.

- These typically come in the form of aligned documents.

- A sentence alignment, using pre-existing document boundaries, is performed automatically. Low-scoring or non-one-to-one sentence alignments are discarded. The resulting aligned sentences constitute the training bitext.

- For many modern SMT systems, induction of word alignments between aligned sentences, using algorithms based on the IBM word-based translation models, is one of the first stages of processing. Such induced word alignments are generally treated as part of the observed data and are used to extract aligned phrases or subtrees.
Modeling

What Makes a Good Translation?
Modeling

• Translation models
  – “Adequacy”
  – Assign better scores to accurate (and complete) translations

• Language models
  – “Fluency”
  – Assign better scores to natural target language text
I did not unfortunately receive an answer to this question.
Blue word links aren’t observed in data.

I did not unfortunately receive an answer to this question.
I did not unfortunately receive an answer to this question.

Blue word links aren’t observed in data.

Features for word-word links: lexica, part-of-speech, orthography, etc.
Word Translation Models

• Usually directed: each word in the target generated by one word in the source
• Many-many and null-many links allowed
• Classic IBM models of Brown et al.
• Used now mostly for word alignment, not translation

Im Anfang war das Wort
In the beginning was the word
I did not unfortunately receive an answer to this question
Auf diese Frage habe ich leider keine Antwort bekommen.

I did not unfortunately receive an answer to this question.
I did not unfortunately receive an answer to this question.
Phrase Translation Models

Not necessarily syntactic phrases
Division into phrases is hidden

Auf diese Frage habe ich leider keine Antwort bekommen

I did not unfortunately receive an answer to this question

Score each phrase pair using several features
Phrase Translation Models

Not necessarily syntactic phrases

Division into phrases is hidden

Auf diese Frage habe ich leider keine Antwort bekommen

I did not unfortunately receive an answer to this question

Score each phrase pair using several features

What are some other features?
Phrase Translation Models

- Capture translations **in context**
  - *en Amerique*: **to** America
  - *en anglais*: **in** English
- State-of-the-art for several years
- Each source/target phrase pair is scored by several weighted features.
- The weighted sum of model features is the whole translation’s score.
- Phrases don’t overlap (cf. language models) but have “reordering” features.
Finite State Models

Kumar, Deng & Byrne, 2005

Source Phrase Segmentation

Source Phrase Reordering

Target Phrase Insertion

Phrase Transduction

Target Phrase Segmentation

Source Language Sentence

grain exports are projected to fall by 25 %

Source Phrases

grain exports are_projected_to fall by_25_%

Reordered Source Phrases

exports grain are_projected_to fall by_25_%

Placement of Target Phrase Insertion Markers

1 exports 1 grain are_projected_to fall by_25_%

Target Phrases

les exportations de grains doivent fléchir de 25 %

Target Language Sentence

les exportations de grains doivent fléchir de 25 %

Kumar, Deng & Byrne, 2005
Finite State Models

First transducer in the pipeline

Here a unigram model of phrases

Map distinct words to phrases

Kumar, Deng & Byrne, 2005
Finite State Models

• Natural composition with other finite state processes, e.g. Chinese word segmentation
• Standard algorithms and widely available tools (e.g. AT&T fsm toolkit)
• Limit reordering to finite offset
• Often impractical to compose all finite state machines offline
Single-Tree Translation Models

I did not unfortunately receive an answer to this question

Parse trees with deeper structure have also been used.
Single-Tree Translation Models

• Either source or target has a hidden tree/parse structure
  – Also known as “tree-to-string” or “tree-transducer” models
• The side with the tree generates words/phrases in tree, not string, order.
• Nodes in the tree also generate words/phrases on the other side.
• English side is often parsed, whether it’s source or target, since English parsing is more advanced.
I did not unfortunately receive an answer to this question.
Tree-Tree Translation Models

• Both sides have hidden tree structure
  – Can be represented with a “synchronous” grammar
• Some models assume isomorphic trees, where parent-child relations are preserved; others do not.
• Trees can be fixed in advance by monolingual parsers or induced from data (e.g. Hiero).
• Cheap trees: project from one side to the other
Latent Seq-Seq Models

- Various methods for building source representation
  - Recurrent NN, LSTM, ConvNN, Neural attention
- Representation replicated at each output position
  - Integrated LM, or combined in beam search
Learning Word Translations from Parallel Text

The “IBM Models”
Lexical translation

• How to translate a word → look up in dictionary

  **Haus** — *house, building, home, household, shell.*

• *Multiple translations*

  – some more frequent than others
  – for instance: *house*, and *building* most common
  – special cases: *Haus* of a *snail* is its *shell*

• Note: During all the lectures, we will translate from a foreign language into English
Collect statistics

• Look at a parallel corpus (German text along with English translation)

<table>
<thead>
<tr>
<th>Translation of <em>Haus</em></th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>house</em></td>
<td>8,000</td>
</tr>
<tr>
<td><em>building</em></td>
<td>1,600</td>
</tr>
<tr>
<td><em>home</em></td>
<td>200</td>
</tr>
<tr>
<td><em>household</em></td>
<td>150</td>
</tr>
<tr>
<td><em>shell</em></td>
<td>50</td>
</tr>
</tbody>
</table>
Estimate translation probabilities

- *Maximum likelihood estimation*

\[
p_f(e) = \begin{cases} 
0.8 & \text{if } e = \text{house}, \\
0.16 & \text{if } e = \text{building}, \\
0.02 & \text{if } e = \text{home}, \\
0.015 & \text{if } e = \text{household}, \\
0.005 & \text{if } e = \text{shell}.
\end{cases}
\]
Alignment

• In a parallel text (or when we translate), we **align** words in one language with the words in the other

```
1   2   3   4
das  Haus  ist  klein
the  house  is  small
```

• Word *positions* are numbered 1–4
Alignment function

• Formalizing *alignment* with an *alignment function*

• Mapping an English target word at position $i$ to a German source word at position $j$ with a function $a : i \rightarrow j$

• Example

\[ a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\} \]
Reordering

- Words may be reordered during translation

\[
a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}
\]
One-to-many translation

- A source word may translate into **multiple** target words

$\begin{align*}
\text{1} & \quad \text{2} & \quad \text{3} & \quad \text{4} \\
\text{das} & \quad \text{Haus} & \quad \text{ist} & \quad \text{klitzeklein} \\
\text{the} & \quad \text{house} & \quad \text{is} & \quad \text{very} & \quad \text{small} \\
\text{1} & \quad \text{2} & \quad \text{3} & \quad \text{4} & \quad \text{5} \\
a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\} \end{align*}$
Dropping words

- Words may be dropped when translated
  - The German article *das* is dropped

```
1 2 3 4
das Haus ist klein
```

```
house is small
```

\(\alpha : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}\)
Inserting words

• Words may be **added** during translation
  
  – The English *just* does not have an equivalent in German
  – We still need to map it to something: special **NULL** token

\[
\begin{align*}
0 & \quad \text{NULL} \\
1 & \quad \text{das} \\
2 & \quad \text{Haus} \\
3 & \quad \text{ist} \\
4 & \quad \text{klein} \\
1 & \quad \text{the} \\
2 & \quad \text{house} \\
3 & \quad \text{is} \\
4 & \quad \text{just} \\
5 & \quad \text{small} \\
\end{align*}
\]

\[a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}\]
IBM Model 1

- **Generative model**: break up translation process into smaller steps
  - IBM Model 1 only uses *lexical translation*

- Translation probability
  - for a foreign sentence \( f = (f_1, \ldots, f_{l_f}) \) of length \( l_f \)
  - to an English sentence \( e = (e_1, \ldots, e_{l_e}) \) of length \( l_e \)
  - with an alignment of each English word \( e_j \) to a foreign word \( f_i \) according to the alignment function \( a : j \rightarrow i \)

  \[
p(e, a | f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{a(j)})
  \]

  - parameter \( \epsilon \) is a *normalization constant*
Example

<table>
<thead>
<tr>
<th>das</th>
<th>Haus</th>
<th>ist</th>
<th>klein</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>e</strong></td>
<td>**t(e</td>
<td>f)**</td>
<td><strong>e</strong></td>
</tr>
<tr>
<td>the</td>
<td>0.7</td>
<td>house</td>
<td>0.8</td>
</tr>
<tr>
<td>that</td>
<td>0.15</td>
<td>building</td>
<td>0.16</td>
</tr>
<tr>
<td>which</td>
<td>0.075</td>
<td>home</td>
<td>0.02</td>
</tr>
<tr>
<td>who</td>
<td>0.05</td>
<td>household</td>
<td>0.015</td>
</tr>
<tr>
<td>this</td>
<td>0.025</td>
<td>shell</td>
<td>0.005</td>
</tr>
</tbody>
</table>

\[
p(e, a|f) = \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein})
\]

\[
= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4
\]

\[
= 0.0028 \epsilon
\]
Learning lexical translation models

• We would like to estimate the lexical translation probabilities $t(e|f)$ from a parallel corpus

• ... but we do not have the alignments

• **Chicken and egg problem**
  
  – if we had the *alignments*,
    
    → we could estimate the *parameters* of our generative model
  
  – if we had the *parameters*,
    
    → we could estimate the *alignments*
EM algorithm

- **Incomplete data**
  - if we had *complete data*, we could estimate *model*
  - if we had *model*, we could fill in the *gaps in the data*

- **Expectation Maximization (EM) in a nutshell**
  - initialize model parameters (e.g. uniform)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate
EM algorithm

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
- Model learns that, e.g., *la* is often aligned with *the*
EM algorithm

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- After one iteration

- Alignments, e.g., between *la* and *the* are more likely
EM algorithm

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

• After another iteration

• It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)
EM algorithm

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

• Convergence

• Inherent hidden structure revealed by EM
EM algorithm

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

\[
\begin{align*}
p(la|\text{the}) &= 0.453 \\
p(le|\text{the}) &= 0.334 \\
p(\text{maison}|\text{house}) &= 0.876 \\
p(\text{bleu}|\text{blue}) &= 0.563
\end{align*}
\]

- Parameter estimation from the aligned corpus
IBM Model 1 and EM

- EM Algorithm consists of two steps

  - **Expectation-Step**: Apply model to the data
    - parts of the model are hidden (here: alignments)
    - using the model, assign probabilities to possible values

  - **Maximization-Step**: Estimate model from data
    - take assign values as fact
    - collect counts (weighted by probabilities)
    - estimate model from counts

- Iterate these steps until convergence
IBM Model 1 and EM

• We need to be able to compute:
  – Expectation-Step: probability of alignments
  – Maximization-Step: count collection
IBM Model 1 and EM

- **Probabilities**

  \[
  p(\text{the}|\text{la}) = 0.7 \quad p(\text{house}|\text{la}) = 0.05 \\
  p(\text{the}|\text{maison}) = 0.1 \quad p(\text{house}|\text{maison}) = 0.8
  \]

- **Alignments**

  \[
  \text{la • the} \quad \text{la • the} \quad \text{la • the} \quad \text{la • the} \\
  \text{maison • house} \quad \text{maison • house} \quad \text{maison • house} \quad \text{maison • house}
  \]
IBM Model 1 and EM

- **Probabilities**
  
  \[ p(\text{the}|\text{la}) = 0.7 \quad p(\text{house}|\text{la}) = 0.05 \]
  
  \[ p(\text{the}|\text{maison}) = 0.1 \quad p(\text{house}|\text{maison}) = 0.8 \]

- **Alignments**

  \[ p(e, a|f) = 0.56 \quad p(e, a|f) = 0.035 \quad p(e, a|f) = 0.08 \quad p(e, a|f) = 0.005 \]
IBM Model 1 and EM

- **Probabilities**
  
  \[
  p(\text{the}|\text{la}) = 0.7 \quad p(\text{house}|\text{la}) = 0.05 \\
  p(\text{the}|\text{maison}) = 0.1 \quad p(\text{house}|\text{maison}) = 0.8
  \]

- **Alignments**

  \[
  \text{la} \rightarrow \text{the} \quad \text{maison} \rightarrow \text{house} \quad \text{la} \rightarrow \text{the} \quad \text{maison} \rightarrow \text{house} \quad \text{la} \rightarrow \text{the} \quad \text{maison} \rightarrow \text{house} \quad \text{la} \rightarrow \text{the} \quad \text{maison} \rightarrow \text{house}
  \]

  \[
  p(\text{e}, \text{a}|\text{f}) = 0.56 \quad p(\text{e}, \text{a}|\text{f}) = 0.035 \quad p(\text{e}, \text{a}|\text{f}) = 0.08 \quad p(\text{e}, \text{a}|\text{f}) = 0.005 \\
  p(\text{a}|\text{e}, \text{f}) = 0.824 \quad p(\text{a}|\text{e}, \text{f}) = 0.052 \quad p(\text{a}|\text{e}, \text{f}) = 0.118 \quad p(\text{a}|\text{e}, \text{f}) = 0.007
  \]

- **Counts**

  \[
  c(\text{the}|\text{la}) = 0.824 + 0.052 \quad c(\text{house}|\text{la}) = 0.052 + 0.007 \\
  c(\text{the}|\text{maison}) = 0.118 + 0.007 \quad c(\text{house}|\text{maison}) = 0.824 + 0.118
  \]
IBM Model 1 and EM: Expectation Step

• We need to compute $p(a|e, f)$

• Applying the chain rule:

$$p(a|e, f) = \frac{p(e, a|f)}{p(e|f)}$$

• We already have the formula for $p(e, a|f)$ (definition of Model 1)
IBM Model 1 and EM: Expectation Step

- We need to compute $p(e|f)$

$$p(e|f) = \sum_{a} p(e, a|f)$$

$$= \sum_{a(1)=0}^{l_f} \ldots \sum_{a(l_e)=0}^{l_f} p(e, a|f)$$

$$= \sum_{a(1)=0}^{l_f} \ldots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$
IBM Model 1 and EM: Expectation Step

\[ p(e|f) = \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_a(j)) \]

\[ = \frac{\epsilon}{(l_f + 1)^{l_e}} \sum_{a(1)=0}^{l_f} \cdots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t(e_j|f_a(j)) \]

\[ = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i) \]

- Note the trick in the last line
  - removes the need for an exponential number of products
  - this makes IBM Model 1 estimation tractable

Philipp Koehn
JHU SS
6 July 2006
IBM Model 1 and EM: Expectation Step

- Combine what we have:

\[
p(a|e,f) = \frac{p(e,a|f)}{p(e|f)}
\]

\[
= \frac{\frac{e}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_a(j))}{\frac{e}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)}
\]

\[
= \prod_{j=1}^{l_e} \frac{t(e_j|f_a(j))}{\sum_{i=0}^{l_f} t(e_j|f_i)}
\]
IBM Model 1 and EM: Maximization Step

- Now we have to *collect counts*

- Evidence from a sentence pair $e, f$ that word $e$ is a translation of word $f$:

\[
c(e|f; e, f) = \sum_{a} p(a|e, f) \sum_{j=1}^{l_e} \delta(e, e_j) \delta(f, f_{a(j)})
\]

- With the same simplification as before:

\[
c(e|f; e, f) = \frac{t(e|f)}{\sum_{j=1}^{l_e} t(e|f_{a(j)})} \sum_{j=1}^{l_e} \delta(e, e_j) \sum_{i=0}^{l_f} \delta(f, f_i)
\]
IBM Model 1 and EM: Maximization Step

• After collecting these counts over a corpus, we can estimate the model:

\[ t(e|f; e, f) = \frac{\sum_{(e,f)} c(e|f; e, f))}{\sum_f \sum_{(e,f)} c(e|f; e, f))} \]
IBM Model 1 and EM: Pseudocode

initialize \( t(e|f) \) uniformly

do
  set count(e|f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs \( (e_s,f_s) \)
    for all words \( e \) in \( e_s \)
      total_s = 0
      for all words \( f \) in \( f_s \)
        total_s += \( t(e|f) \)
      for all words \( e \) in \( e_s \)
        for all words \( f \) in \( f_s \)
          count(e|f) += \( t(e|f) / total_s \)
          total(f) += \( t(e|f) / total_s \)
    for all \( f \) in domain( total(.) )
    for all \( e \) in domain( count(.|f) )
      \( t(e|f) = count(e|f) / total(f) \)
  until convergence
Higher IBM Models

<table>
<thead>
<tr>
<th>IBM Model 1</th>
<th>lexical translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 2</td>
<td>adds absolute reordering model</td>
</tr>
<tr>
<td>IBM Model 3</td>
<td>adds fertility model</td>
</tr>
<tr>
<td>IBM Model 4</td>
<td>relative reordering model</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>fixes deficiency</td>
</tr>
</tbody>
</table>

- Only IBM Model 1 has *global maximum*
  - training of a higher IBM model builds on previous model

- Computationally biggest change in Model 3
  - trick to simplify estimation does not work anymore
  - *exhaustive* count collection becomes computationally too expensive
  - *sampling* over high probability alignments is used instead
IBM Model 4

Mary did not slap the green witch
Mary not slap slap slap the green witch
Mary not slap slap slap NULL the green witch
Maria no daba una bofetada a la bruja verde

n(3|slap)
p-null
t(la|the)
d(4|4)
Word alignment

- Notion of word alignment valuable

- Shared task at NAACL 2003 and ACL 2005 workshops
Word alignment with IBM models

• IBM Models create a \textit{many-to-one} mapping
  
  – words are aligned using an \textbf{alignment function}
  – a function may return the same value for different input (one-to-many mapping)
  – a function can not return multiple values for one input (\textit{no many-to-one} mapping)

• But we need \textit{many-to-many} mappings
Symmetrizing word alignments

- Intersection of GIZA++ bidirectional alignments

Philipp Koehn  JHU SS  6 July 2006
Symmetrizing word alignments

- Grow additional alignment points [Och and Ney, CompLing2003]
Growing heuristic

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ( (-1,0), (0,-1), (1,0), (0,1), (-1,-1), (-1,1), (1,-1), (1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e )
               add alignment point ( e-new, f-new )

FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Philipp Koehn
JHU SS
6 July 2006
Synchronous Grammars: Inversion Transduction Grammar
Syntactically-Motivated Distortion

The Authority will be accountable to the Financial Secretary.
管理局將會向財政司負責。

(Authority will to Financial Secretary accountable.)
Syntactically-Motivated Distortion
ITG Overview

- Special case of synchronous CFG
- One, joint nonterminal per bilingual node
- Children are translated monotonically, or reversed
- Binarized normal form
- Mostly used for exact, polytime alignment
ITG Rules

S → [SP Stop]
SP → [NP VP] | [NP VV] | [NP V]
PP → [Prep NP]
NP → [Det NN] | [Det N] | [Pro] | [NP Conj NP]
NN → [A N] | [NN PP]
VP → [Aux VP] | [Aux VV] | [VV PP]
VV → [V NP] | [Cop A]
Det → the/ε
Prep → to/向
Pro → I/我 | you/你
N → authority/管理局 | secretary/司
A → accountable/負貴 | financial/財政
Conj → and/和
Aux → will/將會
Cop → be/ε
Stop → ./。

VP → (VV PP)
ITG Alignment

Where is the Secretary of Finance when needed?

財政司有需要時在哪裡?
Legal ITG Alignments
Bracketing ITG

A \overset{a}{\rightarrow} [A A]

A \overset{a}{\rightarrow} \langle A A \rangle

A \overset{b_{ij}}{\rightarrow} u_i/v_j \quad \text{for all } i, j \text{ English-Chinese lexical translations}

A \overset{b_{ic}}{\rightarrow} u_i/c \quad \text{for all } i \text{ English vocabulary}

A \overset{b_{ej}}{\rightarrow} \epsilon/v_j \quad \text{for all } j \text{ Chinese vocabulary}
Removing Spurious Ambiguity

A $\rightarrow^a [A\ B]$
A $\rightarrow^a [B\ B]$
A $\rightarrow^a [C\ B]$
A $\rightarrow^a [A\ C]$
A $\rightarrow^a [B\ C]$
B $\rightarrow^a \langle A\ A\ \rangle$
B $\rightarrow^a \langle B\ A\ \rangle$
B $\rightarrow^a \langle C\ A\ \rangle$
B $\rightarrow^a \langle A\ C\ \rangle$
B $\rightarrow^a \langle B\ C\ \rangle$

C $\rightarrow_{b_{ij}} u_i/v_j$  for all $i, j$ English-Chinese lexical translations

C $\rightarrow_{b_{ie}} u_i/\epsilon$  for all $i$ English vocabulary

C $\rightarrow_{b_{ej}} \epsilon/v_j$  for all $j$ Chinese vocabulary
Specialized Translation Models: Named Entities
Translating Words in a Sentence

- Models will automatically learn entries in probabilistic translation dictionaries, for instance $p(\text{elle}|\text{she})$, from co-occurrences in aligned sentences of a parallel text.

- For some kinds of words/phrases, this is less effective. For example:
  - numbers
  - dates
  - named entities (NE)

The reason: these constitute a large open class of words that will not all occur even in the largest bitext. Plus, there are regularities in translation of numbers/dates/NE.
Handling Named Entities

For many language pairs, and particularly those which do not share an alphabet, transliteration of person and place names is the desired method of translation.

General Method:
1. Identify NE’s via classifier
2. Transliterate name
3. Translate/reorder honorifics

Also useful for alignment. Consider the case of Inuktitut-English alignment, where Inuktitut renderings of European names are highly nondeterministic.
Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically.
Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically.

Train a **probabilistic finite-state transducer** to model this ambiguous transformation.
Transliteration

Inuktitut rendering of English names changes the string significantly but not deterministically...

... Mr. Williams ... ... mista uialims ...
Useful Types of Word Analysis

- Number/Date Handling

- Named Entity Tagging/Transliteration

- Morphological Analysis
  - Analyze a word to its root form (at least for word alignment)
    - was -> is
    - believing -> believe
    - ruminerai -> ruminer
    - ruminiez -> ruminer
  - As a dimensionality reduction technique
  - To allow lookup in existing dictionary
Learning Word Translation Dictionaries Using Minimal Resources
Learning Translation Lexicons for Low-Resource Languages

| {Serbian Uzbek Romanian Bengali} | English |

**Problem:** Scarce resources . . .
- Large parallel texts are very helpful, but often unavailable
- Often, no “seed” translation lexicon is available
- Neither are resources such as parsers, taggers, thesauri

**Solution:** Use only monolingual corpora in source, target languages
- But use many information sources to propose and rank translation candidates
Bridge Languages

ENGLISH

Dictionary

Intra-family string transduction
* Constructing translation candidate sets

- **Serbian** → **Czech** → **English**

  - **source word**
  - **bridge words**
  - **translation candidates**

  **granica**
  - Via String Similarity

  **granit**
  - Via Bilingual Dictionary

  **hranice**

  **kronika**
  - (1)
  - memorial

  **(9)**
  - record

  **(10)**
  - granite

  **(11)**
  - pale
  - boundary
  - blow
Cognate Selection

The Bridge Language Paradigm

Tasks

some cognates

Spanish-Italian  homogenizar omogeneizzare
Polish-Serbian  befszyk biftek
German-Dutch  gefestigt gevestigd

<table>
<thead>
<tr>
<th>Spanish Word</th>
<th>Italian Word</th>
<th>Cognate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>electron</td>
<td>elettrone</td>
<td>No</td>
</tr>
<tr>
<td>aventurero</td>
<td>avventuriero</td>
<td>No</td>
</tr>
<tr>
<td>perifrasis</td>
<td>perifrasi</td>
<td>No</td>
</tr>
<tr>
<td>divulgar</td>
<td>divulgare</td>
<td>No</td>
</tr>
<tr>
<td>triada</td>
<td>triade</td>
<td>No</td>
</tr>
<tr>
<td>agresivo</td>
<td>aggressivo</td>
<td>No</td>
</tr>
<tr>
<td>insertar</td>
<td>inserto</td>
<td>No</td>
</tr>
<tr>
<td>esprint</td>
<td>sprint</td>
<td>No</td>
</tr>
<tr>
<td>trópico</td>
<td>tropico</td>
<td>No</td>
</tr>
<tr>
<td>altimetro</td>
<td>altimetro</td>
<td>No</td>
</tr>
<tr>
<td>alegato</td>
<td>lista</td>
<td>No</td>
</tr>
<tr>
<td>variado</td>
<td>variato</td>
<td>No</td>
</tr>
<tr>
<td>cepillar</td>
<td>piallare</td>
<td>No</td>
</tr>
<tr>
<td>confusin</td>
<td>confusione</td>
<td>No</td>
</tr>
<tr>
<td>fortificacion</td>
<td>fortificazione</td>
<td>No</td>
</tr>
<tr>
<td>conjuncion</td>
<td>congiunzione</td>
<td>No</td>
</tr>
<tr>
<td>encantador</td>
<td>incantatore</td>
<td>No</td>
</tr>
<tr>
<td>heredero</td>
<td>erede</td>
<td>No</td>
</tr>
<tr>
<td>vidrio</td>
<td>vetro</td>
<td>No</td>
</tr>
<tr>
<td>vaciar</td>
<td>variare</td>
<td>No</td>
</tr>
<tr>
<td>talisman</td>
<td>talismano</td>
<td>No</td>
</tr>
<tr>
<td>sólido</td>
<td>solido</td>
<td>No</td>
</tr>
<tr>
<td>criptografia</td>
<td>crittografia</td>
<td>No</td>
</tr>
<tr>
<td>carencia</td>
<td>carenza</td>
<td>No</td>
</tr>
<tr>
<td>cortesania</td>
<td>cortesia</td>
<td>No</td>
</tr>
<tr>
<td>sadico</td>
<td>sadico</td>
<td>No</td>
</tr>
<tr>
<td>concentracion</td>
<td>concentrazione</td>
<td>No</td>
</tr>
<tr>
<td>venida</td>
<td>venuta</td>
<td>No</td>
</tr>
<tr>
<td>agonizante</td>
<td>agonizzante</td>
<td>No</td>
</tr>
<tr>
<td>extinguir</td>
<td>estinguere</td>
<td>No</td>
</tr>
</tbody>
</table>
## The Transliteration Problem

### Tasks

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piedade</td>
<td>BEH YEH YEH DAL ALEF DAL YEH</td>
</tr>
<tr>
<td>Bolivia</td>
<td>BEH WAW LAM YEH FEH YEH ALEF</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>LAM KAF SEEN MEEM BEH WAW REH GHAIN</td>
</tr>
<tr>
<td>Zanzibar</td>
<td>ZAIN NOON JEEM YEH BEH ALEF REH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inuktitut</th>
<th>Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Williams</strong>:</td>
<td>uialims uilialums uiliammass viliams</td>
</tr>
<tr>
<td><strong>Campbell</strong>:</td>
<td>kaampu kaampul kamvul kaamvul</td>
</tr>
<tr>
<td><strong>McLean</strong>:</td>
<td>makalain maklainn makliin makkalain</td>
</tr>
</tbody>
</table>
Example Models for Cognate and Transliteration Matching

Memoryless Transducer

(Ristad & Yianilos 1997)

| (a,a) | 0.30 |
| (p,p) | 0.25 |
| (p,b) | 0.15 |
| (_,a) | 0.10 |
| (a,_) | 0.08 |
| (_,b) | 0.07 |
| (p,_) | 0.05 |
Two-State Transducer ("Weak Memory")

Example Models for Cognate and Transliteration Matching
Example Models for Cognate and Transliteration Matching

Unigram Interlingua Transducer
Examples: Possible Cognates Ranked by Various String Models

<table>
<thead>
<tr>
<th>String Transduction Models Ranking Spanish Bridge Words for Romanian Source Word <em>inghiti</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C1</strong></td>
</tr>
</tbody>
</table>

- **Romanian**: *inghiti* (ingest)
- **Uzbek**: *avvalgi* (previous/former)

* Effectiveness of cognate models
* Multi-family bridge languages
Similarity Measures
for re-ranking cognate/transliteration hypotheses

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties
Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties
**Compare Vectors**

<table>
<thead>
<tr>
<th>nezavisnost vector</th>
<th>independence vector</th>
<th>freedom vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection of context vector from Serbian to English term space</td>
<td>Construction of context term vector</td>
<td>Construction of context term vector</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>expression</th>
<th>religion</th>
<th>justice</th>
<th>majesty</th>
<th>sovereignty</th>
<th>declaration</th>
<th>country</th>
<th>ornamental</th>
</tr>
</thead>
<tbody>
<tr>
<td>nezavisnost</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>4</td>
<td>1.5</td>
</tr>
<tr>
<td>independence</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>479</td>
<td>836</td>
<td>191</td>
<td>0</td>
</tr>
<tr>
<td>freedom</td>
<td>681</td>
<td>184</td>
<td>104</td>
<td>0</td>
<td>21</td>
<td>4</td>
<td>141</td>
<td>0</td>
</tr>
</tbody>
</table>

Compute cosine similarity between `nezavisnost` and “independence”

... and between `nezavisnost` and “freedom”
Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties
Date Distribution Similarity

• Topical words associated with real-world events appear within news articles in bursts following the date of the event

• Synonymous topical words in different languages, then, display similar distributions across dates in news text: this can be measured

• We use cosine similarity on date term vectors, with term values $p(\text{word} | \text{date})$, to quantify this notion of similarity
Date Distribution Similarity - Example

DATE (200-Day Window)

\[ p(\text{word}|\text{date}) \]

\[ \text{(correct)} \quad \text{independence} \]

\[ \text{(incorrect)} \quad \text{freedom} \]
Similarity Measures

1. Probabilistic string transducers

2. Context similarity

3. Date distribution similarity

4. Similarities based on monolingual word properties
Relative Frequency

\[
rf(w_F) = \frac{f_{C_F}(w_F)}{|C_F|}
\]

\[
rf(w_E) = \frac{f_{C_E}(w_E)}{|C_E|}
\]

Cross-Language Comparison:

\[
\min\left(\frac{rf(w_F)}{rf(w_E)}, \frac{rf(w_E)}{rf(w_F)}\right)
\]

[min-ratio method]

Precedent in Yarowsky & Wicentowski (2000); used relative frequency similarity for morphological analysis
## Combining Similarities: Uzbek

### Individual Bridge Language Results For Uzbek

<table>
<thead>
<tr>
<th>Rank</th>
<th>Turkish</th>
<th>Russian</th>
<th>Farsi</th>
<th>Kyrgyz</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04</td>
<td>0.12</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>0.23</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>0.13</td>
<td>0.26</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>20</td>
<td>0.16</td>
<td>0.28</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>50</td>
<td>0.21</td>
<td>0.30</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>100</td>
<td>0.24</td>
<td>0.31</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>200</td>
<td>0.26</td>
<td>0.32</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>

### Multiple Bridge Language Results For Uzbek

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tur+Rus</th>
<th>Tur+Rus+Farsi</th>
<th>Tur+Rus+Eng</th>
<th>Tur+Rus+Farsi+Kaz+Kyr</th>
<th>Tur+Rus+Farsi+Kaz+Kyr+Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.12</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>0.26</td>
<td>0.27</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>20</td>
<td>0.35</td>
<td>0.37</td>
<td>0.35</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>50</td>
<td>0.39</td>
<td>0.41</td>
<td>0.39</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>100</td>
<td>0.41</td>
<td>0.43</td>
<td>0.41</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td>200</td>
<td>0.43</td>
<td>0.45</td>
<td>0.42</td>
<td>0.48</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Combining Similarities: Romanian, Serbian, & Bengali

**Multiple Bridge Language Results for Romanian Using Combined Similarity Measures**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Spanish</th>
<th>Spanish + Russian</th>
<th>Spanish + English</th>
<th>Spanish + Russian + English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>0.18</td>
<td><strong>0.19</strong></td>
<td><strong>0.19</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.31</td>
<td>0.35</td>
<td>0.34</td>
<td><strong>0.37</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.37</td>
<td>0.41</td>
<td>0.41</td>
<td><strong>0.43</strong></td>
</tr>
<tr>
<td>20</td>
<td>0.43</td>
<td>0.46</td>
<td>0.46</td>
<td><strong>0.48</strong></td>
</tr>
<tr>
<td>50</td>
<td>0.51</td>
<td>0.53</td>
<td>0.53</td>
<td><strong>0.55</strong></td>
</tr>
<tr>
<td>100</td>
<td>0.57</td>
<td>0.60</td>
<td>0.58</td>
<td><strong>0.61</strong></td>
</tr>
<tr>
<td>200</td>
<td>0.60</td>
<td><strong>0.62</strong></td>
<td>0.59</td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>

**Multiple Bridge Language Results for Serbian Using Combined Similarity Measures**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cz</th>
<th>Rus</th>
<th>Bulg</th>
<th>Cz + English</th>
<th>Cz + Slovak</th>
<th>Cz + Slovak + Rus</th>
<th>Cz + Slovak + Rus + Bulg + English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13</td>
<td>0.15</td>
<td><strong>0.19</strong></td>
<td>0.13</td>
<td><strong>0.19</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.19</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>0.24</td>
<td>0.31</td>
<td>0.25</td>
<td><strong>0.38</strong></td>
<td><strong>0.38</strong></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.29</td>
<td>0.28</td>
<td>0.35</td>
<td>0.30</td>
<td>0.42</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.32</td>
<td>0.31</td>
<td>0.40</td>
<td>0.34</td>
<td>0.48</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.38</td>
<td>0.36</td>
<td>0.44</td>
<td>0.39</td>
<td>0.54</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.40</td>
<td>0.40</td>
<td>0.48</td>
<td>0.42</td>
<td><strong>0.59</strong></td>
<td><strong>0.59</strong></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.41</td>
<td>0.42</td>
<td>0.50</td>
<td>0.43</td>
<td><strong>0.60</strong></td>
<td><strong>0.60</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Bridge Language Results for Bengali Using Combined Similarity Measures**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hindi</th>
<th>Hindi + English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.11</td>
<td><strong>0.14</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.13</td>
<td><strong>0.17</strong></td>
</tr>
<tr>
<td>20</td>
<td>0.16</td>
<td><strong>0.21</strong></td>
</tr>
<tr>
<td>50</td>
<td>0.19</td>
<td><strong>0.25</strong></td>
</tr>
<tr>
<td>100</td>
<td>0.22</td>
<td><strong>0.28</strong></td>
</tr>
<tr>
<td>200</td>
<td>0.23</td>
<td><strong>0.29</strong></td>
</tr>
</tbody>
</table>
Observations

* With no Uzbek-specific supervision, we can produce an Uzbek-English dictionary which is 14% exact-match correct

* Or, we can put a correct translation in the top-10 list 34% of the time (useful for end-to-end machine translation or cross-language information retrieval)

* Adding more bridge languages helps

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tur+Rus</th>
<th>Tur+Rus +Farsi</th>
<th>Tur+Rus +Eng</th>
<th>Tur+Rus +Farsi +Kaz+Kyr</th>
<th>Tur+Rus +Farsi +Kaz+Kyr+Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.12</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>0.26</td>
<td>0.27</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>20</td>
<td>0.35</td>
<td>0.37</td>
<td>0.35</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>50</td>
<td>0.39</td>
<td>0.41</td>
<td>0.39</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td>100</td>
<td>0.41</td>
<td>0.43</td>
<td>0.41</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td>200</td>
<td>0.43</td>
<td>0.45</td>
<td>0.42</td>
<td>0.48</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Topic Models
Jobless rate at 3-year low as payrolls surge

WASHINGTON (Reuters) - The United States created jobs at the fastest pace in nine months in January and the unemployment rate unexpectedly dropped to a near three-year low, giving a boost to President Barack Obama.

Nonfarm payrolls jumped 243,000, the Labor Department said on Friday, as factory jobs grew by the most in a year. The jobless rate fell to 8.3 percent - the lowest since February 2009 - from 8.5 percent in December.

The gain in employment was the largest since April and it far outstripped the 150,000 predicted in a Reuters poll of economists. It hinted at underlying economic strength and lessened chances of further stimulus from the Federal Reserve.

"More pistons in the economic engine have begun to fire, pointing to accelerating economic growth. One of the happiest persons reading this job report is President Obama," said Sung Won Sohn, an economics professor at California State University Channel Islands.

The payroll gains were widespread - from retail to temporary help, and from construction to manufacturing - an indication the recovery was becoming more durable.
Topical Similarity
Genehmigung des Protokolls


Gibt es Einwände?

Die Punkte 3 und 4 widersprechen sich jetzt, obwohl es bei der Abstimmung anders aussah.

Das muß ich erst einmal klären, Frau Oomen-Ruijten.

Approval of the minutes

The minutes of the sitting of Thursday, 28 March 1996 have been distributed.

Are there any comments?

Points 3 and 4 now contradict one another whereas the voting showed otherwise.

I will have to look into that, Mrs Oomen-Ruijten.

Koehn (2005): European Parliament corpus
Multilingual Topical Similarity

Abraham Lincoln

From Wikipedia, the free encyclopedia

This article is about the American president. For other uses, see Abraham Lincoln (disambiguation).

Abraham Lincoln (/əˈbrɛəm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. He successfully led his country through a great constitutional, military and moral crisis – the American Civil War – preserving the Union, while ending slavery, and promoting economic and financial modernization. Reared in a poor family on the western frontier, Lincoln was mostly self-educated. He became a country lawyer, an Illinois state legislator, and a one-term member of the United States House of Representatives, but failed in two attempts to be elected to the United States Senate.

Abraham Lincoln

Seine Präsidentschaft gilt als eine der bedeutendsten in der Geschichte der Vereinigten Staaten: Die Wahl des Sklavereigegners veranlasste zunächst sieben, später weitere vier der sklavenhaltenden Südstaaten zur Sezession. Lincoln führte die verbliebenen Nordstaaten durch den daraus entstandenen Bürgerkrieg, setzte die Wiederherstellung der Union durch und betrieb erfolgreich die Abschaffung der Sklaverei in den Vereinigten Staaten. Unter seiner Regierung schlugen die USA den Weg zum zentral regierten, modernen Industriestaat ein und schufen so die Basis für ihren Aufstieg zur Weltmacht im 20. Jahrhundert.
What Representation?
What Representation?

- Bag of words, n-grams, etc.?
What Representation?

- Bag of words, n-grams, etc.?
- Vocabulary mismatch within language:
What Representation?

- Bag of words, n-grams, etc.?
- Vocabulary mismatch within language:
  - *Jobless vs. unemployed*
What Representation?

• Bag of words, n-grams, etc.?
  • Vocabulary mismatch within language:
    • *Jobless* vs. *unemployed*
  • What about between languages?
What Representation?

- Bag of words, n-grams, etc.?
- Vocabulary mismatch within language:
  - Jobless vs. unemployed
- What about between languages?
  - Translate everything into English?
What Representation?

• Bag of words, n-grams, etc.?
  • Vocabulary mismatch within language:
    • *Jobless* vs. *unemployed*
  • What about between languages?
    • Translate everything into English?
• Represent documents/passages as probability distributions over hidden “topics”
Plate Notation

Figure 2.5: Example graphical model representation of a joint probability distribution $P(X, Y)$ and its plate notation equivalent.

Figure 2.6: Nested plate notation and its unrolled graphical model equivalent.
Plate Notation

Figure 2.5: Example graphical model representation of a joint probability distribution $P(X, Y)$ and its plate notation equivalent.

Figure 2.6: Nested plate notation and its unrolled graphical model equivalent.
Modeling Text with Naive Bayes

- Let the text talk about \( T \) topics
- Each topic is a probability dist’n over all words
- For \( D \) documents each with \( N_D \) words:
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

• Let the text talk about $T$ topics
Modeling Text with Topics

*Latent Dirichlet Allocation* (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
Modeling Text with Topics

*Latent Dirichlet Allocation* (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist'n over all words
- For $D$ documents each with $N_D$ words:

![Diagram showing Latent Dirichlet Allocation]
Modeling Text with Topics

*Latent Dirichlet Allocation* (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
- For $D$ documents each with $N_D$ words:

\[
\begin{align*}
\text{Prior} & \xrightarrow{} \theta \\
\text{Prior} & \xrightarrow{} \phi
\end{align*}
\]
Modeling Text with Topics

*Latent Dirichlet Allocation* (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
- For $D$ documents each with $N_D$ words:

80% economy
20% pres. elect.
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
- For $D$ documents each with $N_D$ words:

$\theta \rightarrow z \rightarrow \phi$

Prior

80% economy
20% pres. elect.
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
- For $D$ documents each with $N_D$ words:

$$
\text{Prior} \rightarrow \theta \rightarrow z \rightarrow \phi \rightarrow \text{Prior}
$$

80% economy
20% pres. elect.

econom
Modeling Text with Topics

*Latent Dirichlet Allocation* (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability distribution over all words
- For $D$ documents each with $N_D$ words:

![Diagram of Latent Dirichlet Allocation](image)

80% economy
20% pres. elect.

econom
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

- Let the text talk about $T$ topics
- Each topic is a probability dist’n over all words
- For $D$ documents each with $N_D$ words:

80% economy
20% pres. elect.

80% economy
20% pres. elect.
Figure 2.8: Unrolled graphical model representation of LDA.
Multinomials as Histograms

Figure 2.2: Examples of three multinomial distributions over ten topics. The y-axis represents the probability of the topic being present in the document.

\[
P_{k}^{i} = \frac{c_{i}}{Q_{k}^{i}}
\]

\[
P_{k}^{i} = \frac{c_{i} + 1}{Q_{k}^{i} + \sum_{i=1}^{K} c_{i}}
\]

Figure 2.2 shows examples of three multinomial distributions over ten possible outcomes which in this case are equivalent to a set of ten possible topics that have been assigned to three documents.

2.2.1.4 Dirichlet Distribution

The Dirichlet distribution is a continuous distribution over a family of multinomial distributions. Often abbreviated as "Dir", this distribution is parametrized by a vector \( \alpha \): Dir(\( \alpha \)), referred as the hyperparameter, that controls the sparsity of the family of multinomial distributions. In a \( K \)-dimensional probability simplex the pdf of the Dirichlet distribution is defined as:

\[
p(x_{1}, x_{2}, x_{3}, \ldots, x_{n} | \alpha_{1}, \alpha_{2}, \alpha_{3}, \ldots, \alpha_{n}) = \frac{\prod_{i=1}^{K} \binom{c_{i} + \alpha_{i} - 1}{c_{i}}}{\prod_{i=1}^{K} \binom{\sum_{i=1}^{K} c_{i} + \alpha_{i} - 1}{\sum_{i=1}^{K} c_{i}}}
\]

\[
(2.7)
\]
Dirichlet Priors on Histograms

Figure 2.3: Example of multinomial distribution samples drawn from Dirichlet distributions with different symmetric hyperparameters (i.e. concentration parameters).

\( \alpha = 0.01 \)

\( \alpha = 0.1 \)

\( \alpha = 1.0 \)

\( \alpha = 10.0 \)

\( \alpha = 100.0 \)
Top Words by Topic

Topics →

1. DISEASE
   BACTERIA
   DISEASES
   GERMS
   FEVER
   CAUSE
   CAUSED
   SPREAD
   VIRUSES
   INFECTION
   VIRUS
   MICROORGANISMS
   PERSON
   INFECTIOUS
   COMMON
   CAUSING
   SMALLPOX
   BODY
   INFECTIONS
   CERTAIN

2. WATER
   FISH
   SEA
   SWIM
   SWIMMING
   POOL
   LIKE
   SHELL
   SHARK
   TANK
   SHELLS
   SHARKS
   DIVING
   DOLPHINS
   SWAM
   LONG
   SEAL
   DIVE
   DOLPHIN
   UNDERWATER

3. MIND
   WORLD
   DREAM
   DREAMS
   THOUGHT
   THOUGHTS
   OWN
   REAL
   LIFE
   IMAGINE
   SENSE
   CONSCIOUSNESS
   STRANGE
   FEELING
   WHOLE
   BEING
   MIGHT
   HOPE

4. STORY
   STORIES
   TELL
   CHARACTERS
   AUTHOR
   READ
   TOLD
   SETTING
   TALES
   PLOT
   TELLING
   SHORT
   FICTION
   ACTION
   TRUE
   EVENTS
   TELLS
   TALE
   NOVEL

5. FIELD
   MAGNETIC
   MAGNET
   WIRE
   NEEDLE
   CURRENT
   COIL
   POLES
   IRON
   COMPASS
   LINES
   CORE
   ELECTRIC
   DIRECTION
   FORCE
   MAGNETS
   BE
   MAGNETISM
   POLE
   INDUCED

6. SCIENCE
   STUDY
   SCIENTISTS
   SCIENTIFIC
   KNOWLEDGE
   WORK
   RESEARCH
   CHEMISTRY
   TECHNOLOGY
   MANY
   MATHEMATICS
   BIOLOGY
   PHYSICS
   LABORATORY
   STUDIES
   WORLD
   SCIENTIST
   STUDYING
   SCIENCES

7. BALL
   GAME
   TEAM
   BASEBALL
   PLAYERS
   PLAY
   FIELD
   PLAYER
   BASKETBALL
   COACH
   PLAYED
   PLAYING
   HIT
   TENNIS
   TEAMS
   GAMES
   SPORTS
   BAT
   TERRY

8. JOB
   WORK
   JOBS
   CAREER
   EXPERIENCE
   EMPLOYMENT
   OPPORTUNITIES
   WORKING
   TRAINING
   SKILLS
   CAREERS
   POSITIONS
   FIND
   POSITION
   FIELD
   OCCUPATIONS
   REQUIRE
   OPPORTUNITY
   EARN
   ABLE
### Top Words by Topic

**Topics →**

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISEASE</td>
<td>WATER</td>
<td>MIND</td>
<td>STORY</td>
<td>FIELD</td>
<td>SCIENCE</td>
<td>BALL</td>
<td>JOB</td>
</tr>
<tr>
<td>BACTERIA</td>
<td>FISH</td>
<td>WORLD</td>
<td>STORIES</td>
<td>MAGNETIC</td>
<td>STUDY</td>
<td>GAME</td>
<td>WORK</td>
</tr>
<tr>
<td>DISEASES</td>
<td>SEA</td>
<td>DREAM</td>
<td>TELL</td>
<td>MAGNET</td>
<td>SCIENTISTS</td>
<td>TEAM</td>
<td>JOBS</td>
</tr>
<tr>
<td>GERMS</td>
<td>SWIM</td>
<td>DREAMS</td>
<td>CHARACTER</td>
<td>WIRE</td>
<td>KNOWLEDGE</td>
<td>FOOTBALL</td>
<td>CAREER</td>
</tr>
<tr>
<td>FEVER</td>
<td>SWIMMING</td>
<td>THOUGHT</td>
<td>AUTHOR</td>
<td>NEEDLE</td>
<td>WORK</td>
<td>BASEBALL</td>
<td>EXPERIENCE</td>
</tr>
<tr>
<td>CAUSE</td>
<td>POOL</td>
<td>IMAGINATION</td>
<td>READ</td>
<td>CURRENT</td>
<td>RESEARCH</td>
<td>PLAY</td>
<td>EMPIRE</td>
</tr>
<tr>
<td>CAUSED</td>
<td>LIKE</td>
<td>MOMENT</td>
<td>TOLD</td>
<td>COIL</td>
<td>CHEMISTRY</td>
<td>FIELD</td>
<td>OPPORTUNITIES</td>
</tr>
<tr>
<td>SPREAD</td>
<td>SHELL</td>
<td>THOUGHTS</td>
<td>SETTING</td>
<td>POLES</td>
<td>TECHNOLOGY</td>
<td>BASKETBALL</td>
<td>WORKING</td>
</tr>
<tr>
<td>VIRUSES</td>
<td>SHARK</td>
<td>OWN</td>
<td>TALES</td>
<td>IRON</td>
<td>MANY</td>
<td>PLAYER</td>
<td>TRAINING</td>
</tr>
<tr>
<td>INFECTION</td>
<td>TANK</td>
<td>REAL</td>
<td>PLOT</td>
<td>COMPASS</td>
<td>MATHEMATICS</td>
<td>FIELD</td>
<td>SKILLS</td>
</tr>
<tr>
<td>VIRUS</td>
<td>SHELLS</td>
<td>LIFE</td>
<td>TELLING</td>
<td>LINES</td>
<td>BIOLOGY</td>
<td>PHYSICS</td>
<td>CAREERS</td>
</tr>
<tr>
<td>MICROORGANISMS</td>
<td>SHARKS</td>
<td>IMAGINE</td>
<td>SHORT</td>
<td>CORE</td>
<td>MANY</td>
<td>LABORATORY</td>
<td>POSITIONS</td>
</tr>
<tr>
<td>PERSON</td>
<td>DIVING</td>
<td>SENSE</td>
<td>FICTION</td>
<td>ELECTRIC</td>
<td>FIELD</td>
<td>STUDIES</td>
<td>FIND</td>
</tr>
<tr>
<td>INFECTIONOUS</td>
<td>DOLPHINS</td>
<td>CONSCIOUSNESS</td>
<td>ACTION</td>
<td>DIRECTION</td>
<td>PHYSICS</td>
<td>WORLD</td>
<td>POSITION</td>
</tr>
<tr>
<td>COMMON</td>
<td>SWAM</td>
<td>STRANGE</td>
<td>TRUE</td>
<td>FORCE</td>
<td>LABORATORY</td>
<td>SCIENTIST</td>
<td>FIELD</td>
</tr>
<tr>
<td>CAUSING</td>
<td>LONG</td>
<td>FEELING</td>
<td>EVENTS</td>
<td>MAGNETS</td>
<td>STUDIES</td>
<td>STUDYING</td>
<td>OCCUPATIONS</td>
</tr>
<tr>
<td>SMALLPOX</td>
<td>SEAL</td>
<td>WHOLE</td>
<td>TELLS</td>
<td>BE</td>
<td>SCIENCES</td>
<td>SCIENCES</td>
<td>REQUIRE</td>
</tr>
<tr>
<td>BODY</td>
<td>DIVE</td>
<td>BEING</td>
<td>TALE</td>
<td>MAGNETISM</td>
<td></td>
<td></td>
<td>OPPORTUNITY</td>
</tr>
<tr>
<td>INFECTIONS</td>
<td>DOLPHIN</td>
<td>MIGHT</td>
<td>NOVEL</td>
<td>POLE</td>
<td></td>
<td></td>
<td>EARN</td>
</tr>
<tr>
<td>CERTAIN</td>
<td>UNDERWATER</td>
<td>HOPE</td>
<td></td>
<td>INDUCED</td>
<td></td>
<td></td>
<td>ABLE</td>
</tr>
</tbody>
</table>

*Griffiths et al.*
Hierarchical Document Models

Example mlhLDA representation of an Astrophysical Journal article

1. INTRODUCTION
Blazars are an intriguing class of active galactic nuclei (AGNs), dominated by non-thermal radiation over the entire electromagnetic spectrum. Their emission extends from radio to TeV energies with a broadband spectral energy distribution (SED) typically described by two main components, the first peaking from IR to X-ray energy range in which blazars are the most commonly detected extragalactic sources...

7. SUMMARY AND DISCUSSION
We have presented the infrared characterization of a sample of blazars detected in the γ-ray. In order to perform our selection, we considered all the blazars in the ROMA-BZCAT catalog (Massaro et al. 2010) that are associated with a γ-ray source in the 2FGL (The Fermi-LAT Collaboration 2011). Then, we searched for infrared counterparts in the WISE archive adopting the same criteria described...

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic = 32</th>
<th>Topic = 48</th>
<th>Topic = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>spectral</td>
<td>measured</td>
<td>aperture</td>
</tr>
<tr>
<td>2</td>
<td>amplification</td>
<td>uncertainties</td>
<td>measured</td>
</tr>
<tr>
<td>3</td>
<td>isotropic</td>
<td>catalog</td>
<td>total</td>
</tr>
<tr>
<td>4</td>
<td>dropout</td>
<td>matching</td>
<td>exposure</td>
</tr>
<tr>
<td>5</td>
<td>competition</td>
<td>estimated</td>
<td>position</td>
</tr>
<tr>
<td>6</td>
<td>caustic</td>
<td>respectively</td>
<td>ratio</td>
</tr>
<tr>
<td>7</td>
<td>detected</td>
<td>final</td>
<td>selected</td>
</tr>
<tr>
<td>8</td>
<td>antenna</td>
<td>cathode</td>
<td>color</td>
</tr>
<tr>
<td>9</td>
<td>function</td>
<td>total</td>
<td>spread</td>
</tr>
<tr>
<td>10</td>
<td>color</td>
<td>limit</td>
<td>objects</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic = 20</th>
<th>Topic = 48</th>
<th>Topic = 90</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>entanglement</td>
<td>measured</td>
<td>ferroelectric</td>
</tr>
<tr>
<td>2</td>
<td>color</td>
<td>uncertainties</td>
<td>population</td>
</tr>
<tr>
<td>3</td>
<td>distance</td>
<td>catalog</td>
<td>rational</td>
</tr>
<tr>
<td>4</td>
<td>magnitude</td>
<td>matching</td>
<td>fraction</td>
</tr>
<tr>
<td>5</td>
<td>accretion</td>
<td>estimated</td>
<td>starburst</td>
</tr>
<tr>
<td>6</td>
<td>similar</td>
<td>respectively</td>
<td>shielding</td>
</tr>
<tr>
<td>7</td>
<td>modulus</td>
<td>final</td>
<td>similar</td>
</tr>
<tr>
<td>8</td>
<td>objects</td>
<td>cathode</td>
<td>emitting</td>
</tr>
<tr>
<td>9</td>
<td>right</td>
<td>total</td>
<td>reputation</td>
</tr>
<tr>
<td>10</td>
<td>parameters</td>
<td>limit</td>
<td>respectively</td>
</tr>
</tbody>
</table>
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

Multiple languages?
Modeling Text with Topics

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003)

Multiple languages?
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Topic: set of distributions on words
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document tuples: text in 1 or more languages

Topic: set of distributions on words
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document tuples: text in 1 or more languages

Topic: set of distributions on words

Scales linearly w/ number of langs. (unlike pairwise)
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document tuples: text in 1 or more languages

Compare text in different languages with $\theta$, i.e. topic distribution

Topic: set of distributions on words

Scales linearly w/ number of langs. (unlike pairwise)
Multilingual Text with Topics

Polylingual Topic Models (EMNLP 2009)

Document *tuples*: text in 1 or more languages

Compare text in different languages with $\theta$, i.e. topic distribution

Topic: set of distributions on words

Scales linearly w/ number of langs. (unlike pairwise)

But...

- No phrase translations
- No distinction of parallel, comparable text
- No modeling of document features (e.g., length)
Genehmigung des Protokolls

Gibt es Einwände?
Die Punkte 3 und 4 widersprechen sich jetzt, obwohl es bei der Abstimmung anders aussah.

Das muß ich erst einmal klären, Frau Oomen-Ruijten.

Approval of the minutes
The minutes of the sitting of Thursday, 28 March 1996 have been distributed.

Are there any comments?
Points 3 and 4 now contradict one another whereas the voting showed otherwise.

I will have to look into that, Mrs Oomen-Ruijten.

Koehn (2005): European Parliament corpus
Example Europarl Topics

DA centralbank europæiske ecb s lån centralbanks
DE zentralbank ezb bank europäischen investitionsbank darlehen
EL τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN bank central ecb banks european monetary
ES banco central europeo bce bancos centrales
FI keskuspankin ekp n euroopan keskuspankki eip
FR banque centrale bce européenne banques monétaire
IT banca centrale bce europea banche prestiti
NL bank centrale ecb europese banken leningen
PT banco central europeu bce bancos empréstimos
SV centralbanken europeiska ecb centralbankens s lån

\[ T = 400 \]
Example Europarl Topics

### Objectives

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>mål nå målsætninger målet målsætning opnå</td>
</tr>
<tr>
<td>DE</td>
<td>ziel ziele erreichen zielen erreicht zielsetzungen</td>
</tr>
<tr>
<td>EL</td>
<td>στόχους στόχο στόχος στόχων στόχοι επίτευξη</td>
</tr>
<tr>
<td>EN</td>
<td>objective objectives achieve aim ambitious set</td>
</tr>
<tr>
<td>ES</td>
<td>objetivo objetivos alcanzar conseguir lograr estos</td>
</tr>
<tr>
<td>FI</td>
<td>tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen</td>
</tr>
<tr>
<td>FR</td>
<td>objectif objectifs atteindre but cet ambitieux</td>
</tr>
<tr>
<td>IT</td>
<td>obiettivo obiettivi raggiungere degli scopo quello</td>
</tr>
<tr>
<td>NL</td>
<td>doelstellingen doel doelstelling bereiken bereikt doelen</td>
</tr>
<tr>
<td>PT</td>
<td>objectivo objectivos alcançar atingir ambicioso conseguir</td>
</tr>
<tr>
<td>SV</td>
<td>mål målet uppnå målen målsättningar målsättning</td>
</tr>
</tbody>
</table>

$$T = 400$$
### Example Europarl Topics

<table>
<thead>
<tr>
<th>Code</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>andre anden side ene andet ørige</td>
</tr>
<tr>
<td>DE</td>
<td>anderen andere einen wie andererseits anderer</td>
</tr>
<tr>
<td>EL</td>
<td>álles álla álλη álλων álλους óπως</td>
</tr>
<tr>
<td>EN</td>
<td>other one hand others another there</td>
</tr>
<tr>
<td>ES</td>
<td>otros otras otro otra parte demás</td>
</tr>
<tr>
<td>FI</td>
<td>muiden toisaalta muita muut muihin muun</td>
</tr>
<tr>
<td>FR</td>
<td>autres autre part côté ailleurs même</td>
</tr>
<tr>
<td>IT</td>
<td>altri altre altro altra dall parte</td>
</tr>
<tr>
<td>NL</td>
<td>andere anderzijds anderen ander als kant</td>
</tr>
<tr>
<td>PT</td>
<td>outros outras outro lado outra noutros</td>
</tr>
<tr>
<td>SV</td>
<td>andra sidan å annat ena annan</td>
</tr>
</tbody>
</table>

\[ T = 400 \]
Multilingual Topical Similarity
Example Wikipedia Topics

\[ T = 400 \]

CY  sadwrn blaned gallair at lloeren mytholeg
DE  space nasa sojus flug mission
EL  διαστημικό sts nasa αγγλ small
EN  space mission launch satellite nasa spacecraft
FA  فضایی ماموریت ناسا مدار فضانورد ماهواره
FI  sojuz nasa apollo ensimmäinen space lento
FR  spatiale mission orbite mars satellite spatial
HE  התלול האזר חלול כדור א ת conosc
IT  spaziale missione programma space sojuz stazione
PL  misja kosmicznej stacji misji space nasa
RU  космический союз космического спутник станции
TR  uzay soyuz ay uzaya salyut sovyetler
Example Wikipedia Topics

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY</td>
<td>sbaen madrid el la josé sbaeneg</td>
</tr>
<tr>
<td>DE</td>
<td>de spanischer spanischen spanien madrid la</td>
</tr>
<tr>
<td>EL</td>
<td>ισπανίας ισπανία de ισπανός ντε μαδρίτη</td>
</tr>
<tr>
<td>EN</td>
<td>de spanish spain la madrid y</td>
</tr>
<tr>
<td>FA</td>
<td>ترين de اسبانيا اسبانيايي كوبا مادريد</td>
</tr>
<tr>
<td>FI</td>
<td>espanja de espanjan madrid la real</td>
</tr>
<tr>
<td>FR</td>
<td>espagnol espagne madrid espagnole juan y</td>
</tr>
<tr>
<td>HE</td>
<td>ספרד ספרדית דה مدريد הספרדית קובא</td>
</tr>
<tr>
<td>IT</td>
<td>de spagna spagnolo spagnola madrid el</td>
</tr>
<tr>
<td>PL</td>
<td>de hiszpański hiszpanii la juan y</td>
</tr>
<tr>
<td>RU</td>
<td>де мадрид испании испания испанский de</td>
</tr>
<tr>
<td>TR</td>
<td>ispanya ispanyol madrid la küba real</td>
</tr>
</tbody>
</table>

$T = 400$
Example Wikipedia Topics

\[ T = 400 \]
Differences in Topic Emphasis
Differences in Topic Emphasis

world ski km won
Differences in Topic Emphasis

world ski km won

actor role television actress
Differences in Topic Emphasis

- World ski km won
- Ottoman empire khan byzantine
- Actor role television actress
Document Inference

Latent Dirichlet Allocation (LDA)

Polylingual Topic Model (PLTM)
Document Inference

Latent Dirichlet Allocation (LDA)

Polylingual Topic Model (PLTM)

Inference
Extracted English-Spanish news stories from the Gigaword collection using PLTM trained on OCD output:

**EN:** WASHINGTON, URGENT: Treasury chief defends dollar as world reserve currency. US Treasury Secretary Timothy Geithner said Wednesday that "the dollar remains the world's standard reserve currency", following China's call for a new global currency as an alternative to the greenback.

**ES:** WASHINGTON, URGENTE: Washington quiere que el dólar se mantenga como la principal divisa de reserva. El secretario del Tesoro estadounidense Timothy Geithner declaró este miércoles que el dólar se mantiene como la principal moneda mundial de reserva y que Estados Unidos bregará porque se mantenga como tal.

He(EN,ES)=0.055

**ES:** BUENOS AIRES: Peso argentino estable a 3,70 por dólar. La moneda argentina se mantuvo estable este miércoles a 3,70 pesos por dólar, según el promedio de bancos y casas de cambio. El Banco Central viene interviniendo en el mercado para administrar una devaluación gradual de la moneda con respecto al dólar estadounidense.

He(EN,ES)=0.153

**ES:** Washington: EEUU quiere que el dólar se mantenga como la principal divisa de reserva. El secretario del Tesoro estadounidense Timothy Geithner declaró este miércoles que el dólar se mantiene como la principal moneda mundial de reserva y que Estados Unidos bregará porque se mantenga como tal. "Pienso que el dólar sigue siendo la moneda de reserva de referencia y pienso que debería continuar siéndolo durante largo tiempo", declaró Geithner ante el Consejo de Relaciones Exteriores en Nueva York. "Como país haremos lo necesario para conservar la confianza en nuestros mercados financieros" y en nuestra economía, agregó.

He(EN,ES)=0.086

**ES:** WASHINGTON: Obama defiende derecho a la expansión de la OTAN. El presidente estadounidense Barack Obama dijo este miércoles que Estados Unidos quería "reiniciar" las relaciones con Rusia pero añadió que la OTAN debería de todos modos estar abierta a los países que aspiren a unirse a esa alianza. "Mi gobierno busca reiniciar las relaciones con Rusia", dijo Obama al cabo de una reunión en la Casa Blanca con el secretario general de la OTAN, Jaap de Hoop Scheffer. Pero dijo que los renovados vínculos con Moscú deben ser "consistentes con la membresía de la OTAN y consistentes con la necesidad de enviar una clara señal en Europa de que vamos a atenernos (...)."

He(EN,ES)=0.172
Training MT from Comparable Corpora

- **MT system performance - parallel vs. extracted sentences**
  - Parallel collection: News Commentary (all) & Europarl (all)
  - Extracted Sentences: Gigaword (4 years)

<table>
<thead>
<tr>
<th>Training Source</th>
<th>Collection Size</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parallel</td>
<td>Extracted</td>
</tr>
<tr>
<td>News Commentary (NC)</td>
<td>131K</td>
<td>0</td>
</tr>
<tr>
<td>Europarl (EU)</td>
<td>1.75M</td>
<td>0</td>
</tr>
<tr>
<td>Gigaword Extracted (GE)</td>
<td>0</td>
<td>926K</td>
</tr>
<tr>
<td>NC+GE</td>
<td>131K</td>
<td>926K</td>
</tr>
<tr>
<td>EU+GE</td>
<td>1.75M</td>
<td>926K</td>
</tr>
</tbody>
</table>

- **News (WMT’11)**: 23.75, 23.91, 24.25, 24.92, 25.90
- **Europarl (WMT’09)**: 25.43, 32.06, 23.88, 25.61, 31.59

Krstovski, 2016
Bilingual Embeddings

Figure 2: Forms of supervision required by the four models compared in this paper. From left to right, the cost of the supervision required varies from expensive (BiSkip) to cheap (BiVCD). BiSkip requires a parallel corpus annotated with word alignments (Fig. 2a), BiCVM requires a sentence-aligned corpus (Fig. 2b), BiCCA only requires a bilingual lexicon (Fig. 2c) and BiVCD requires comparable documents (Fig. 2d).

*Upadhyay et al. (2016)*
Bilingual Embeddings

Figure 3: PCA projection of word embeddings of some frequent words present in English-French corpus. English and French words are shown in blue and red respectively.

<table>
<thead>
<tr>
<th>Language</th>
<th>Mono</th>
<th>BiSkip</th>
<th>BiCVM</th>
<th>BiCCA</th>
<th>BiVCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>71.1</td>
<td>72.0</td>
<td>60.4</td>
<td>71.4</td>
<td>58.9</td>
</tr>
<tr>
<td>fr</td>
<td>78.9</td>
<td>80.4</td>
<td>73.7</td>
<td>80.2</td>
<td>69.5</td>
</tr>
<tr>
<td>sv</td>
<td>75.5</td>
<td>78.2</td>
<td>70.5</td>
<td>79.0</td>
<td>64.5</td>
</tr>
<tr>
<td>zh</td>
<td>73.8</td>
<td>73.1</td>
<td>65.8</td>
<td>71.7</td>
<td>67.0</td>
</tr>
<tr>
<td>avg.</td>
<td>74.8</td>
<td>75.9</td>
<td>67.6</td>
<td>75.6</td>
<td>66.8</td>
</tr>
</tbody>
</table>

Table 7: Labeled attachment score (LAS) for dependency parsing when trained and tested on language $l$. Mono refers to parser trained with mono-lingually induced embeddings. Scores in bold are better than the Mono scores for each language, showing improvement from cross-lingual training.

An interesting observation is that BiCCA and BiVCD are better at separating antonyms. The words peace and war, (and their French translations paix and guerre) are well separated in BiCCA and BiVCD. However, in BiSkip and BiCVM these pairs are very close together. This can be attributed to the fact that BiSkip and BiCVM are trained on parallel sentences, and if two antonyms are present in the same sentence in English, they will also be present together in its French translation. However, BiCCA uses bilingual dictionary and BiVCD use comparable sentence context, which helps in pulling apart the synonyms and antonyms.

6 Discussion

The goal of this paper was to formulate the task of learning cross-lingual word vector representations in a unified framework, and conduct experiments to compare the performance of existing methods.
Search

What’s the best translation (under our model)?
Search

• Even if we know the right words in a translation, there are $n!$ permutations.
  
  $10! = 3,626,800 \quad 20! \approx 2.43 \times 10^{18} \quad 30! \approx 2.65 \times 10^{32}$

• We want the translation that gets the highest score under our model
  – Or the best $k$ translations
  – Or a random sample from the model’s distribution

• But **not** in $n!$ time!
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
Search in Phrase Models

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren

That is why we have

Translate in target language order to ease language modeling.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

That is why we have every reason.

Translate in target language order to ease language modeling.
Search in Phrase Models

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren

That is why we have every reason to

Translate in target language order to ease language modeling.
That is why we have every reason to integrate the Umwelt in the Agrarpolitik.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

That is why we have every reason to integrate the environment.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

That is why we have every reason to integrate the environment in.
That is why we have every reason to integrate the environment in the Agrarpolitik zu integrieren.
That is why we have every reason to integrate the environment in the agricultural policy.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

That is why we have every reason to integrate the environment in the agricultural policy.

Translate in target language order to ease language modeling.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

That is why we have every reason to integrate the environment in the agricultural policy.

Translate in target language order to ease language modeling.
That is why we have every reason to integrate the environment in the agricultural policy.
<table>
<thead>
<tr>
<th>Deshalb</th>
<th>haben</th>
<th>wir</th>
<th>allen</th>
<th>Grund</th>
<th>die</th>
<th>Umwelt</th>
<th>in</th>
<th>die</th>
<th>Agrarpolitik</th>
<th>zu</th>
<th>integrieren</th>
</tr>
</thead>
<tbody>
<tr>
<td>that is why we have</td>
<td>every reason</td>
<td>the environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore</td>
<td>have</td>
<td>we</td>
<td>every reason</td>
<td>the</td>
<td>environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>,</td>
<td>to integrate</td>
<td></td>
</tr>
<tr>
<td>that is why</td>
<td>we have</td>
<td>all</td>
<td>reason</td>
<td>,</td>
<td>which</td>
<td>environment in</td>
<td>agricultural policy</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>have therefore</td>
<td>us</td>
<td>all the</td>
<td>reason</td>
<td>of the</td>
<td>environment into</td>
<td>the agricultural policy</td>
<td>successfully integrated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence</td>
<td>, we</td>
<td>every</td>
<td>reason to make</td>
<td>environmental</td>
<td>on</td>
<td>the cap</td>
<td>be woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>we have therefore</td>
<td>everyone</td>
<td>grounds for taking the</td>
<td>the</td>
<td>environment</td>
<td>to the</td>
<td>agricultural policy is</td>
<td>on</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>, we</td>
<td>all of</td>
<td>cause</td>
<td>which</td>
<td>environment ,</td>
<td>to</td>
<td>the cap ,</td>
<td>for</td>
<td>incorporated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence our</td>
<td>any</td>
<td>why</td>
<td>that</td>
<td>outside</td>
<td>at</td>
<td>agricultural policy</td>
<td>too</td>
<td>woven together</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore ,</td>
<td>it</td>
<td>of all</td>
<td>reason for</td>
<td>, the</td>
<td>completion</td>
<td>into</td>
<td>that agricultural policy</td>
<td>be</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

And many, many more…even before reordering
<table>
<thead>
<tr>
<th>Deshalb</th>
<th>haben</th>
<th>wir</th>
<th>allen</th>
<th>Grund</th>
<th>,</th>
<th>die</th>
<th>Umwelt</th>
<th>in</th>
<th>die</th>
<th>Agrarpolitik</th>
<th>zu</th>
<th>integrieren</th>
</tr>
</thead>
<tbody>
<tr>
<td>that is why we have</td>
<td>every reason</td>
<td>the environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore</td>
<td>have</td>
<td>we</td>
<td>every reason</td>
<td>the</td>
<td>environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that is why</td>
<td>we have</td>
<td>all</td>
<td>reason</td>
<td>,</td>
<td>which</td>
<td>environment in</td>
<td>agricultural policy</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>have therefore</td>
<td>us</td>
<td>all the</td>
<td>reason</td>
<td>of the</td>
<td>environment into</td>
<td>the</td>
<td>agricultural policy</td>
<td>successfully integrated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence</td>
<td>, we</td>
<td>every</td>
<td>reason to make</td>
<td>environmental</td>
<td>on</td>
<td>the cap</td>
<td>be woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>we have therefore</td>
<td>everyone</td>
<td>grounds for taking the</td>
<td>the</td>
<td>environment</td>
<td>to the</td>
<td>agricultural policy</td>
<td>is</td>
<td>on</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>, we</td>
<td>all of</td>
<td>cause</td>
<td>which</td>
<td>environment ,</td>
<td>to</td>
<td>the</td>
<td>cap ,</td>
<td>for</td>
<td>incorporated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence our</td>
<td>any</td>
<td>why</td>
<td>that</td>
<td>outside</td>
<td>at</td>
<td>agricultural policy</td>
<td>too</td>
<td>woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore ,</td>
<td>it</td>
<td>of all</td>
<td>reason for</td>
<td>, the</td>
<td>completion</td>
<td>into</td>
<td>that agricultural policy</td>
<td>be</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*And many, many more…even before reordering*
| Deshalb | haben | wir | allen | Grund | , | die | Umwelt | in | die | Agrarpolitik | zu | integrieren |
|---------|-------|-----|-------|-------| | die | Umwelt | in | die | Agrarpolitik | zu | integriieren |
| that is why we have | every reason | the environment | in | the | agricultural policy | to | integrate |
| therefore | have | we | every reason | the | environment | in | the | agricultural policy | to | integrate |
| that is why | we have | all | reason | , | which | environment in | agricultural policy | parliament |
| have therefore | us | all the | reason | of the | environment into | the agricultural policy | successfully integrated |
| hence | , we | every | reason to make | environmental | on | the cap | be woven together |
| we have therefore | everyone | grounds for taking the | the | environment | to the | agricultural policy | is | on | parliament |
| so | , we | all of | cause | which | environment , | to | the | cap , | for | incorporated |
| hence our | any | why | that | outside | at | agricultural policy | too | woven together |
| therefore , | it | of all | reason for | , the | completion | into | that | agricultural policy | be |
**Search in Phrase Models**

<table>
<thead>
<tr>
<th>Deshalb</th>
<th>haben</th>
<th>wir</th>
<th>allen</th>
<th>Grund</th>
<th>,</th>
<th>die</th>
<th>Umwelt</th>
<th>in</th>
<th>die</th>
<th>Agrarpolitik</th>
<th>zu</th>
<th>integrieren</th>
</tr>
</thead>
<tbody>
<tr>
<td>that is why we have</td>
<td>every reason</td>
<td>the environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore</td>
<td>have</td>
<td>we</td>
<td>every reason</td>
<td>the</td>
<td>environment</td>
<td>in the</td>
<td>agricultural policy</td>
<td>,</td>
<td>to integrate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that is why we have</td>
<td>all</td>
<td>reason</td>
<td>,</td>
<td>which</td>
<td>environment in</td>
<td>agricultural policy</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>have therefore</td>
<td>us</td>
<td>all the</td>
<td>reason</td>
<td>of the</td>
<td>environment into</td>
<td>the agricultural policy</td>
<td>successfully integrated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence</td>
<td>, we</td>
<td>every</td>
<td>reason to make</td>
<td>environmental</td>
<td>on</td>
<td>the cap</td>
<td>be woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>we have therefore</td>
<td>everyone</td>
<td>grounds for taking the</td>
<td>the</td>
<td>environment</td>
<td>to the</td>
<td>agricultural policy</td>
<td>is</td>
<td>on</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>, we</td>
<td>all of</td>
<td>cause</td>
<td>which</td>
<td>environment</td>
<td>, to</td>
<td>the</td>
<td>cap</td>
<td>,</td>
<td>for</td>
<td>incorporated</td>
<td></td>
</tr>
<tr>
<td>hence our</td>
<td>any</td>
<td>why</td>
<td>that</td>
<td>outside</td>
<td>at</td>
<td>agricultural policy</td>
<td>too</td>
<td>woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore ,</td>
<td>it</td>
<td>of all</td>
<td>reason for</td>
<td>,</td>
<td>the</td>
<td>completion</td>
<td>into</td>
<td>that agricultural policy</td>
<td>be</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*And many, many more…even before reordering*
<table>
<thead>
<tr>
<th>Deshalb</th>
<th>haben</th>
<th>wir</th>
<th>allen</th>
<th>Grund</th>
<th>die</th>
<th>Umwelt</th>
<th>in</th>
<th>die</th>
<th>Agrarpolitik</th>
<th>zu</th>
<th>integrieren</th>
</tr>
</thead>
<tbody>
<tr>
<td>that is why</td>
<td>we have</td>
<td>every reason</td>
<td>the environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore</td>
<td>have</td>
<td>we</td>
<td>every reason</td>
<td>the</td>
<td>environment</td>
<td>in</td>
<td>the</td>
<td>agricultural policy</td>
<td>to</td>
<td>integrate</td>
<td></td>
</tr>
<tr>
<td>that is why</td>
<td>we have</td>
<td>all</td>
<td>reason</td>
<td>, which</td>
<td>environment in</td>
<td>agricultural policy</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>have therefore</td>
<td>us</td>
<td>all</td>
<td>the</td>
<td>reason</td>
<td>of the</td>
<td>environment into</td>
<td>the agricultural policy</td>
<td>successfully integrated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hence</td>
<td>, we</td>
<td>every</td>
<td>reason to make</td>
<td>environmental</td>
<td>on</td>
<td>the cap</td>
<td>be woven together</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>we have therefore</td>
<td>everyone</td>
<td>grounds for taking the</td>
<td>the environment</td>
<td>to the</td>
<td>agricultural policy is</td>
<td>on</td>
<td>parliament</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>so</td>
<td>, we</td>
<td>all</td>
<td>of</td>
<td>cause</td>
<td>which</td>
<td>environment ,</td>
<td>to</td>
<td>the cap ,</td>
<td>for</td>
<td>incorporated</td>
<td></td>
</tr>
<tr>
<td>hence our</td>
<td>any</td>
<td>why</td>
<td>that</td>
<td>outside</td>
<td>at</td>
<td>agricultural policy</td>
<td>too</td>
<td>woven together</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>therefore ,</td>
<td>it</td>
<td>of all</td>
<td>reason for</td>
<td>, the</td>
<td>completion</td>
<td>into</td>
<td>that agricultural policy</td>
<td>be</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

And many, many more…even before reordering
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

hence

we
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

hence
we
have
in
the
Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

"Stack Decoding"
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

hence |
hence we |

we |
we have |

have |
in |
the
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

hence __________ hence we
we ________ we have
have ________ we have
in
the
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

hence we have therefore
in the environment
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

"hence__hence we"

"we__we have"

"we have__we have therefore"

"have__we have"

"in__the environment"

"the"
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.

We could declare these equivalent.
“Stack Decoding”

Deshalb haben wir allen Grund, die Umwelt in die Agrarpolitik zu integrieren.
Many ways of segmenting source
Many ways of translating each segment
*Restrict* model class: phrases >, e.g., 7 words, no long-distance reordering
*Recombine* equivalent hypotheses
*Prune* away unpromising partial translations or we’ll run out of space and/or run too long

– How to compare partial translations?
– Some start with easy stuff: “in”, “das”, ...
– Some with hard stuff: “Agrarpolitik”, “Entscheidungsproblem”, …
Hypothesis Recombination

• Different paths to the same partial translation

$p=1$  Mary  $p=0.534$  did not give  $p=0.092$

$\text{did not}$  $p=0.164$  give  $p=0.044$
Hypothesis Recombination

- Different paths to the same partial translation
- Combine paths
  - Drop weaker path
  - Keep backpointer to weaker path (for lattice or n-best generation)
Hypothesis Recombination

• Recombined hypotheses do not have to match completely
• Weaker path can be dropped if
  – Last n target words match (for n+1-gram lang. model)
  – Source coverage vectors match (same best future)
Hypothesis Recombination

• Combining partially matching hypotheses
Pruning

• Hypothesis recombination is not sufficient

  Heuristically discard weak hypotheses early

• Organize Hypothesis in stacks, e.g. by
  – same foreign words covered
  – same number of foreign words covered
  – same number of English words produced

• Compare hypotheses in stacks, discard bad ones
  – histogram pruning: keep top $n$ hypotheses in each stack (e.g., $n=100$)
  – threshold pruning: keep hypotheses that are at most $t$ times the cost of
    best hypothesis in stack (e.g., $t = 0.001$)
Word Lattice Generation

- **Search graph** can be easily converted into a **word lattice**
  - can be further mined for **n-best lists**
    - enables **reranking** approaches
    - enables **discriminative training**
• Organization of hypothesis into stacks
  – here: based on number of foreign words translated
  – during translation all hypotheses from one stack are expanded
  – expanded Hypotheses are placed into stacks
Limits on Reordering

• Reordering may be limited
  – Monotone Translation: No reordering at all
  – Only phrase movements of at most $n$ words

• Reordering limits speed up search (polynomial instead of exponential)

• Current reordering models are weak, so limits improve translation quality
Comparing Hypotheses

- Comparing hypotheses with *same number of foreign words* covered
  
  Maria no dio una bofetada a la bruja verde

  - e: Mary did not
  - f: **-------*
  - p: 0.154

  **better**
  
  partial translation

  - e: the
  - f: -----**--*
  - p: 0.354

  **covers**
  
  easier part
  
  --> lower cost

- Hypothesis that covers *easy part* of sentence is preferred
  
  Need to consider *future cost* of uncovered parts
  
  or: have one hypothesis stack per coverage vector
Synchronous Grammars

• Just like monolingual grammars except...
  – Each rule involves pairs (tuples) of nonterminals
  – Tuples of elementary trees for TAG, etc.
• First proposed for source-source translation in compilers
• Can be constituency, dependency, lexicalized, etc.
• Parsing speedups for monolingual grammar don’t necessarily work
  – E.g., no split-head trick for lexicalized parsing
• Binarization less straightforward
A variant of CKY chart parsing.
Bilingual Parsing

The fox knows many things.
Bilingual Parsing

The fox knows many things.

<table>
<thead>
<tr>
<th></th>
<th>póll’</th>
<th>oíd’</th>
<th>alópēx</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NP/NP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fox</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP/VP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knows</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>many</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP/VP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>things</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bilingual Parsing

The fox knows many things.

Table:

<table>
<thead>
<tr>
<th></th>
<th>póll’</th>
<th>oîd’</th>
<th>alópēx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fox</td>
<td></td>
<td></td>
</tr>
<tr>
<td>know</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>many</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>things</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S/S
MT as Parsing

• If we only have the source, parse it while recording all compatible target language trees.
• Runtime is also multiplied by a grammar constant: one string could be a noun and a verb phrase
• Continuing problem of multiple hidden configurations (trees, instead of phrases) for one translation.
Parsing as Deduction

\[ \forall A, B, C \in N, W \in V, 0 \leq i, j, k \leq m \]

\[ \text{constit}(B, i, j) \land \text{constit}(C, j, k) \land A \rightarrow BC \Rightarrow \text{constit}(A, i, k) \]

\[ \text{word}(W, i) \land A \rightarrow W \Rightarrow \text{constit}(A, i, i + 1) \]

\[ \text{constit}(A, i, k) = \bigvee_{B, C, j} \text{constit}(B, i, j) \land \text{constit}(C, j, k) \land A \rightarrow B \ C \]

\[ \text{constit}(A, i, j) = \bigvee_{W} \text{word}(W, i, j) \land A \rightarrow W \]
Parsing as Deduction

\[ \text{constit}(A, i, k) = \bigvee_{B,C,j} \text{constit}(B, i, j) \land \text{constit}(C, j, k) \land A \rightarrow B \; C \]

\[ \text{constit}(A, i, j) = \bigvee_{W} \text{word}(W, i, j) \land A \rightarrow W \]

\[ \text{score}(\text{constit}(A, i, k)) = \max_{B,C,j} \text{score}(\text{constit}(B, i, j)) \]
\[ \quad \cdot \text{score}(\text{constit}(C, j, k)) \]
\[ \quad \cdot \text{score}(A \rightarrow B \; C) \]

\[ \text{score}(\text{constit}(A, i, j)) = \max_{W} \text{score}(\text{word}(W, i, j)) \cdot \text{score}(A \rightarrow W) \]

And how about the inside algorithm?
Bilingual Parsing: ITG

\[ s(X, i, k, u, w) = \bigvee_{j,v,Y,Z} s(Y, i, j, u, v) \land s(Z, j, k, v, w) \land [X \to Y \ Z] \]

\[ s(X, i, k, u, w) = \bigvee_{j,v,Y,Z} s(Y, i, j, v, w) \land s(Z, j, k, u, v) \land \langle X \to Y \ Z \rangle \]

\[ s(X, i, j, u, v) = w(S, i, j) \land w(T, u, v) \land X \to S/T \]

\[ s(X, i, j, u, u) = w(S, i, j) \land X \to S/\epsilon \]

\[ s(X, i, i, u, v) = w(T, u, v) \land X \to \epsilon/T \]

Similar extensions for finding the best alignment or the expectations of particular alignments
What Makes Search Hard?

• What we really want: the best (highest-scoring) translation
• What we get: the best translation/phrase segmentation/alignment
  – Even summing over all ways of segmenting one translation is hard.
• Most common approaches:
  – Ignore problem
  – Sum over top $j$ translation/segmentation/alignment triples to get top $k<<j$ translations
Redundancy in $n$-best Lists

Source: Da ich wenig Zeit habe, gehe ich sofort in medias res.
Training

Which features of data predict good translations?
Training: Generative/Discriminative

• Generative
  – Maximum likelihood training: max p(data)
  – “Count and normalize”
  – Maximum likelihood with hidden structure
    • Expectation Maximization (EM)

• Discriminative training
  – Maximum conditional likelihood
  – Minimum error/risk training
  – Other criteria: perceptron and max. margin
“Count and Normalize”

• Language modeling example: assume the probability of a word depends only on the previous 2 words.

\[ p(\text{disease} | \text{into the}) = \frac{p(\text{into the disease})}{p(\text{into the})} \]

• \( p(\text{disease}|\text{into the}) = \frac{3}{20} = 0.15 \)

• “Smoothing” reflects a prior belief that \( p(\text{breech}|\text{into the}) > 0 \) despite these 20 examples.
I did not unfortunately receive an answer to this question.

Auf diese Frage habe ich leider keine Antwort bekommen.
Phrase Models

<table>
<thead>
<tr>
<th>I</th>
<th>did</th>
<th>unfortunately</th>
<th>receive</th>
<th>an</th>
<th>answer</th>
<th>to</th>
<th>this</th>
<th>question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auf</td>
<td>diese</td>
<td>Frage</td>
<td>habe</td>
<td>ich</td>
<td>leider</td>
<td>keine</td>
<td>Antwort</td>
<td>bekomen</td>
</tr>
</tbody>
</table>

Some good phrase pairs.
I

I did not unfortunately receive an answer to this question.

Auf diese Frage habe leider keine Antwort bekommen.

Some bad phrase pairs.
“Count and Normalize”

• Usual approach: treat relative frequencies of source phrase $s$ and target phrase $t$ as probabilities

\[
p(s \mid t) \equiv \frac{\text{count}(s, t)}{\text{count}(t)} \quad p(t \mid s) \equiv \frac{\text{count}(s, t)}{\text{count}(s)}
\]

• This leads to overcounting when not all segmentations are legal due to unaligned words.
Hidden Structure

• But really, we don’t observe word alignments.
• How are word alignment model parameters estimated?
• Find (all) structures consistent with observed data.
  – Some links are incompatible with others.
  – We need to score complete sets of links.
Hidden Structure and EM

• Expectation Maximization
  – Initialize model parameters (randomly, by some simpler model, or otherwise)
  – Calculate probabilities of hidden structures
  – Adjust parameters to maximize likelihood of observed data given hidden data
  – Iterate

• Summing over all hidden structures can be expensive
  – Sum over 1-best, k-best, other sampling methods
Discriminative Training

• Given a source sentence, give “good” translations a higher score than “bad” translations.
• We care about good translations, not a high probability of the training data.
• Spend less “energy” modeling bad translations.
• Disadvantages
  – We need to run the translation system at each training step.
  – System is tuned for one task (e.g. translation) and can’t be directly used for others (e.g. alignment)
“Good” Compared to What?

• Compare current translation to
• Idea #1: a human translation. OK, but
  – Good translations can be very dissimilar
  – We’d need to find hidden features (e.g. alignments)
• Idea #2: other top \( n \) translations (the “n-best list”). Better in practice, but
  – Many entries in n-best list are the same apart from hidden links
• Compare with a **loss function** \( L \)
  – 0/1: wrong or right; equal to reference or not
  – Task-specific metrics (word error rate, BLEU, …)
MT Evaluation

* Intrinsic
  
  Human evaluation

  Automatic (machine) evaluation

* Extrinsic

How useful is MT system output for…

Deciding whether a foreign language blog is about politics?
Cross-language information retrieval?
Flagging news stories about terrorist attacks?
...

...
### Human Evaluation

<table>
<thead>
<tr>
<th>Statement</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Je suis fatigué.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tired is I.</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Cookies taste good!</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>I am exhausted.</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Human Evaluation

PRO

High quality

CON

Expensive!

Person (preferably bilingual) must make a time-consuming judgment per system hypothesis.

Expense prohibits frequent evaluation of incremental system modifications.
Automatic Evaluation

PRO

Cheap. Given available reference translations, free thereafter.

CON

We can only measure some proxy for translation quality. (Such as N-Gram overlap or edit distance).
In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.
Partially excellent translations

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.
In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.
Evaluation of Machine Translation Systems

**Bleu** *(Papineni, Roukos, Ward and Zhu, 2002):*

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.
**Unigram Precision**

- **Unigram Precision** of a candidate translation:

\[
\frac{C'}{N}
\]

where \( N \) is number of words in the candidate, \( C' \) is the number of words in the candidate which are in at least one reference translation.

e.g.,

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

\[
Precision = \frac{17}{18}
\]

(only *obeys* is missing from all reference translations)
**Modified Unigram Precision**

- Problem with unigram precision:

  Candidate: the the the the the the the

  Reference 1: the cat sat on the mat

  Reference 2: there is a cat on the mat

  precision = \( \frac{7}{7} = 1 \)???

- **Modified unigram precision:** “Clipping”
  
  - Each word has a “cap”. e.g., \( cap(\text{the}) = 2 \)
  
  - A candidate word \( w \) can only be correct a maximum of \( cap(w) \) times. e.g., in candidate above, \( cap(\text{the}) = 2 \), and \( \text{the} \) is correct twice in the candidate \( \Rightarrow \)

    \[
    \text{Precision} = \frac{2}{7}
    \]
Modified N-gram Precision

- Can generalize modified unigram precision to other n-grams.

- For example, for candidates 1 and 2 above:

\[
\text{Precision}_1(\text{bigram}) = \frac{10}{17}
\]

\[
\text{Precision}_2(\text{bigram}) = \frac{1}{13}
\]
**Precision Alone Isn’t Enough**

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

\[ \text{Precision}(\text{unigram}) = 1 \]

\[ \text{Precision}(\text{bigram}) = 1 \]
But Recall isn’t Useful in this Case

- Standard measure used in addition to precision is recall:

\[ \text{Recall} = \frac{C}{N} \]

where \( C \) is number of n-grams in candidate that are correct, \( N \) is number of words in the references.

Candidate 1: I always invariably perpetually do.

Candidate 2: I always do

Reference 1: I always do

Reference 1: I invariably do

Reference 1: I perpetually do
Sentence Brevity Penalty

- Step 1: for each candidate, compute closest matching reference (in terms of length)
  e.g., our candidate is length 12, references are length 12, 15, 17. Best match is of length 12.

- Step 2: Say $l_i$ is the length of the $i$’th candidate, $r_i$ is length of best match
  for the $i$’th candidate, then compute

  \[
  \text{brevity} = \frac{\sum_i r_i}{\sum_i l_i}
  \]

  (I think! from the Papineni paper, although \( \text{brevity} = \frac{\sum_i r_i}{\min(l_i, r_i)} \) might
  make more sense?)

- Step 3: compute brevity penalty

  \[
  BP = \begin{cases} 
  1 & \text{If } \text{brevity} < 1 \\
  e^{1-\text{brevity}} & \text{If } \text{brevity} \geq 1 
  \end{cases}
  \]

  e.g., if $r_i = 1.1 \times l_i$ for all $i$ (candidates are always 10% too short) then
  \[BP = e^{-0.1} = 0.905\]
The Final Score

- Corpus precision for any n-gram is

\[
P_n = \frac{\sum_{C \in \text{Candidate}} \sum_{ngram \in C} \text{Count}_{\text{clip}}(ngram)}{\sum_{C \in \text{Candidate}} \sum_{ngram \in C} \text{Count}(ngram)}
\]

i.e. number of correct ngrams in the candidates (after “clipping”) divided by total number of ngrams in the candidates

- Final score is then

\[
\text{Bleu} = BP \times (p_1 p_2 p_3 p_4)^{1/4}
\]

i.e., \(BP\) multiplied by the geometric mean of the unigram, bigram, trigram, and four-gram precisions
Automatic Evaluation: **Bleu Score**

**hypothesis 1**  I am exhausted

**hypothesis 2**  Tired is I

**reference 1**  I am tired

**reference 2**  I am ready to sleep now
## Automatic Evaluation: **Bleu Score**

<table>
<thead>
<tr>
<th></th>
<th>I am exhausted</th>
<th>Tired is I</th>
<th>I I I</th>
<th>I am tired</th>
<th>I am ready to sleep now and so exhausted</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypothesis 1</td>
<td>3/3</td>
<td>1/2</td>
<td>0/1</td>
<td>1/3</td>
<td>0/2</td>
</tr>
<tr>
<td>hypothesis 2</td>
<td>1/3</td>
<td>0/2</td>
<td>0/1</td>
<td>1/3</td>
<td>0/2</td>
</tr>
<tr>
<td>hypothesis 3</td>
<td>1/3</td>
<td>0/2</td>
<td>0/1</td>
<td>1/3</td>
<td>0/2</td>
</tr>
<tr>
<td>reference 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reference 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How Good are Automatic Metrics?

slide from G. Doddington (NIST)
Correlation? [Callison-Burch et al., 2006]

- DARPA/NIST MT Eval 2005
  - Mostly statistical systems (all but one in graphs)
  - One submission **manual post-edit** of statistical system’s output
  - Good adequacy/fluency scores *not reflected* by BLEU
Correlation? [Callison-Burch et al., 2006]

- Comparison of
  - *good statistical* system: high BLEU, high adequacy/fluency
  - *bad statistical* sys. (trained on less data): low BLEU, low adequacy/fluency
  - *Systran*: lowest BLEU score, but high adequacy/fluency

[from Callison-Burch et al., 2006, EACL]
How Good are Automatic Metrics?

- Do n-gram methods like BLEU overly favor certain types of systems?
- Automatic metrics still useful
- During development of one system, a better BLEU indicates a better system
- Evaluating different systems has to depend on human judgement
- What are some other evaluation ideas?
Minimizing Error/Maximizing Bleu

• Adjust parameters to minimize error \( L \) when translating a training set

• Error as a function of parameters is
  – nonconvex: not guaranteed to find optimum
  – piecewise constant: slight changes in parameters might not change the output.

• Usual method: optimize one parameter at a time with linear programming
Generative/Discriminative Reunion

• Generative models can be cheap to train: “count and normalize” when nothing’s hidden.
• Discriminative models focus on problem: “get better translations”.
• Popular combination
  – Estimate several generative translation and language models using relative frequencies.
  – Find their optimal (log-linear) combination using discriminative techniques.
Generative/Discriminative Reunion

Score each hypothesis with several generative models:

$$
\theta_1 p_{\text{phrase}}(\tilde{s} | \tilde{t}) + \theta_2 p_{\text{phrase}}(\tilde{t} | \tilde{s}) + \theta_3 p_{\text{lexical}}(s | t) + L + \theta_7 p_{\text{LM}}(\tilde{t}) + \theta_8 \# \text{words} + L
$$

If necessary, renormalize into a probability distribution:

$$
Z = \sum_k \exp(\hat{\theta} \cdot f_k)
$$

where $k$ ranges over all hypotheses. We then have

$$
p(t_i | s) = \frac{1}{Z} \exp(\hat{\theta} \cdot f)
$$

for any given hypothesis $i$. Unnecessary if thetas sum to 1 and p’s are all probabilities. Exponentiation makes it positive.
Minimizing Risk

Instead of the error of the 1-best translation, compute expected error (risk) using \( k \)-best translations; this makes the function differentiable.

Smooth probability estimates using gamma to even out local bumpiness. Gradually increase gamma to approach the 1-best error.

\[
E_{p_{\gamma,\theta}} \left[ L(s, t) \right]
\]

\[
p_{\gamma,\theta}(t_i \mid s_i) = \frac{[\exp\hat{e} \cdot f_i]^{\gamma}}{\sum_{k'} [\exp\hat{e} \cdot f_{k'}]^{\gamma}}
\]
Encoder-Decoder Models

(cf. Socher & Manning 2016)
Language Models

A language model computes a probability for a sequence of words: \( P(w_1, \ldots, w_T) \)

- Useful for machine translation
  - Word ordering:
    \( p(\text{the cat is small}) > p(\text{small the is cat}) \)
  - Word choice:
    \( p(\text{walking home after school}) > p(\text{walking house after school}) \)
Ngram LMs

• Performance improves with keeping around higher n-grams counts and doing smoothing and so-called backoff (e.g. if 4-gram not found, try 3-gram, etc)

• There are A LOT of n-grams!
  → Gigantic RAM requirements!

• Recent state of the art: Scalable Modified Kneser-Ney Language Model Estimation by Heafield et al.:
  “Using one machine with 140 GB RAM for 2.8 days, we built an unpruned model on 126 billion tokens”
Remember Word Embeddings
The vast majority of rule-based and statistical NLP work regards words as atomic symbols: *hotel, conference, walk*

In vector space terms, this is a vector with one 1 and a lot of zeroes

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a “one-hot” representation. Its problem:

\[
\text{motel } \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{bmatrix} \quad \text{AND} \\
\text{hotel } \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} = 0
\]
Distributional Similarity

You can get a lot of value by representing a word by means of its neighbors.

“You shall know a word by the company it keeps”

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into banking crises as has happened in
saying that Europe needs unified banking regulation to replace the hodgepodge

⇒ These words will represent banking

You can vary whether you use local or large context to get a more syntactic or semantic clustering.
Hard/Soft Clustering

Class based models learn word classes of similar words based on distributional information (~ class HMM)

- Brown clustering (Brown et al. 1992)
- Exchange clustering (Martin et al. 1998, Clark 2003)
- Desparsification and great example of unsupervised pre-training

Soft clustering models learn for each cluster/topic a distribution over words of how likely that word is in each cluster

- Latent Semantic Analysis (LSA/LSI), Random projections
- Latent Dirichlet Analysis (LDA), HMM clustering
Distributed Representation

Similar idea

Combine vector space semantics with the prediction of probabilistic models (Bengio et al. 2003, Collobert & Weston 2008, Turian et al. 2010)

In all of these approaches, including deep learning models, a word is represented as a dense vector

\[
\begin{bmatrix}
0.286 \\
0.792 \\
-0.177 \\
-0.107 \\
0.109 \\
-0.542 \\
0.349 \\
0.271
\end{bmatrix}
\]

linguistics =
Visualizing Embeddings
Vector Semantics

Mikolov, Yih & Zweig (2013)

These representations are *way* better at encoding dimensions of similarity than we realized!

- Analogies testing dimensions of similarity can be solved quite well just by doing vector subtraction in the embedding space

  Syntactically
  - \( x_{apple} - x_{apples} \approx x_{car} - x_{cars} \approx x_{family} - x_{families} \)
  - Similarly for verb and adjective morphological forms

Semantically (Semeval 2012 task 2)
  - \( x_{shirt} - x_{clothing} \approx x_{chair} - x_{furniture} \)
Recurrent Neural Nets

- RNNs tie the weights at each time step
- Condition the neural network on all previous words
- RAM requirement only scales with number of words
RNN LMs

Given list of word vectors: \( x_1, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_T \)

At a single time step:

\[
\begin{align*}
    h_t &= \sigma \left( W^{(hh)} h_{t-1} + W^{(hx)} x_{[t]} \right) \\
    \hat{y}_t &= \text{softmax} \left( W^{(S)} h_t \right) \\
    \hat{P}(x_{t+1} = v_j \mid x_t, \ldots, x_1) &= \hat{y}_{t,j}
\end{align*}
\]
RNN LMs

Main idea: we use the same set of W weights at all time steps!

Everything else is the same:
\[
\begin{align*}
h_t &= \sigma \left( W^{(hh)} h_{t-1} + W^{(hx)} x[t] \right) \\
\hat{y}_t &= \text{softmax} \left( W^{(S)} h_t \right) \\
\hat{P}(x_{t+1} = v_j \mid x_t, \ldots, x_1) &= \hat{y}_{t,j}
\end{align*}
\]

$h_0 \in \mathbb{R}^{D_h}$ is some initialization vector for the hidden layer at time step 0

$x[t]$ is the column vector of L at index [t] at time step t

$W^{(hh)} \in \mathbb{R}^{D_h \times D_h}$, $W^{(hx)} \in \mathbb{R}^{D_h \times d}$, $W^{(S)} \in \mathbb{R}^{V \times D_h}$
RNN LMs

\( \hat{y} \in \mathbb{R}^{|V|} \) is a probability distribution over the vocabulary

Same cross entropy loss function but predicting words instead of classes

\[
J^{(t)}(\theta) = - \sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}
\]
Training RNNs is hard!

• Multiply the same matrix at each time step during forward prop

• Ideally inputs from many time steps ago can modify output $y$

• Take $\frac{\partial E_2}{\partial W}$ for an example RNN with 2 time steps! Insightful!
Clipping Gradients

- The solution first introduced by Mikolov is to clip gradients to a maximum value.

```
Algorithm 1 Pseudo-code for norm clipping the gradients whenever they explode

\[ \hat{g} \leftarrow g \frac{||g||}{\text{threshold}} \]
```

- Makes a big difference in RNNs.
Words are assigned to "classes" based on their unigram frequency

First, class layer is evaluated; then, only words belonging to the predicted class are evaluated, instead of the whole output layer $y$

[Goodman2001]

Provides speedup in some cases more than $100\times$
Perplexity Results

KN5 = Count-based language model with Kneser-Ney smoothing & 5-grams

Table from paper *Extensions of recurrent neural network language model* by Mikolov et al 2011
The improvement obtained from a single RNN model over the best backoff model increases with more data!
Deeper

(Irsoy & Cardie, 2014)
RNNs

\[ h_t = f(Wx_t + Vh_{t-1} + b) \]
\[ y_t = g(Uh_t + c) \]

- \( x \) represents a token (word) as a vector.
- \( y \) represents the output label.
- \( h \) is the memory, computed from the past memory and current word. It summarizes the sentence up to that time.
Label Bias

• In some state space configurations, MEMMs (and RNNs) essentially ignore the inputs

• This is not a problem for HMMs and CRFs
Bidirectionality

\[ h_t = f(Wx_t + Vh_{t-1} + b) \]
\[ \overrightarrow{h_t} = f(Wx_t + \overrightarrow{Vh_{t+1}} + \overrightarrow{b}) \]
\[ y_t = g(U[h_t; \overrightarrow{h_t}] + c) \]

\( h = [\overrightarrow{h}; \overleftarrow{h}] \) now represents (summarizes) the past and future around a single token.
Going Deep

Are recurrent networks really deep? (e.g. like this)
Each memory layer passes an intermediate sequential representation to the next.
Opinion Mining

Fine-grained opinion analysis aims to detect subjectivity (e.g. “hate”) and characterize
- Intensity (e.g. strong)
- Sentiment (e.g. negative)
- Opinion holder, target or topic

... 

Important for a variety of NLP tasks such as
- Opinion-oriented question answering
- Opinion summarization
In this work, we focus on detecting *direct subjective expressions* (DSEs) and *expressive subjective expressions* (ESEs).

DSE: Explicit mentions of private states or speech events expressing private states
ESE: Expressions that indicate sentiment, emotion, etc. without explicitly conveying them.
The committee, [as usual]_{ESE}, [has refused to make any statements]_{DSE}.

In BIO notation (where a token is the atomic unit):

```
O O O O B_ESE I_ESE O B_DSE
I_DSE I_DSE I_DSE I_DSE I_DSE O
```
CRF et al.

Success of CRF based approaches hinges critically on access to a good feature set, typically based on
• Constituency and dependency parse trees
• Manually crafted opinion lexicons
• Named entity taggers
• Other preprocessing components
(See Yang and Cardie (2012) for an up-to-date list.)

What about feature learning?
Hypotheses

We expected that deep recurrent nets would improve upon shallow recurrent nets, especially on ESE extraction.
- ESEs are harder to identify: They are variable in length and might involve terms that are neutral in most contexts (e.g. “as usual”, “in fact”).

How the networks would perform against (semi)CRFs was unclear, especially when CRFs are given access to word vectors.
### Results: Examples

<table>
<thead>
<tr>
<th>Method</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>The situation obviously remains fluid from hour to hour but it [seems to be] [going in the right direction]</td>
</tr>
<tr>
<td>Deep RntNN</td>
<td>The situation [obviously] remains fluid from hour to hour but it [seems to be going in the right] direction</td>
</tr>
<tr>
<td>Shallow RntNN</td>
<td>The situation [obviously] remains fluid from hour to hour but it [seems to be going in] the right direction</td>
</tr>
<tr>
<td>Semi-CRF</td>
<td>The situation [obviously remains fluid from hour to hour but it seems to be going in the right direction]</td>
</tr>
</tbody>
</table>