(Mostly Statistical) Machine Translation

Kemal Oflazer
The Rosetta Stone

• Decree from Ptolemy V on repealing taxes and erecting some statues (196 BC)
• Written in three languages
  – Hieroglyphic
  – Demotic
  – Classical Greek
Overview

• History of Machine Translation
• Early Rule-based Approaches
• Introduction to Statistical Machine Translation (SMT)
• Advanced Topics in SMT
• Evaluation of (S)MT output
Machine Translation

• Transform text (speech) in one language (source) to text (speech) in a different language (target) such that
  – The “meaning” in the source language input is (mostly) preserved, and
  – The target language output is grammatical.

• Holy grail application in AI/NLP since middle of 20th century.
Translation

• Process
  – Read the text in the source language
  – **Understand** it
  – **Write** it down in the target language

• These are hard tasks for computers
  – The human process is invisible, intangible
Many possible legitimate translations!
Rolls-Royce Merlin Engine
(from German Wikipedia)

• Der Rolls-Royce Merlin ist ein 12-Zylinder-
  Flugmotor von Rolls-Royce in V-Bauweise,
  der vielen wichtigen britischen und US-
  amerikanischen Flugzeugmustern des
  Zweiten Weltkriegs als Antrieb diente. Ab
  1941 wurde der Motor in Lizenz von der
  Packard Motor Car Company in den USA als
  Packard V-1650 gebaut.

• Nach dem Krieg wurden diverse Passagier-
  und Frachtflugzeuge mit diesem Motor
  ausgestattet, so z. B. Avro Lancastrian, Avro
  Tudor und Avro York, später noch einmal die
  Canadair C-4 (umgebaute Douglas C-54). Der
  zivile Einsatz des Merlin hielt sich jedoch in
  Grenzen, da er als robust, aber zu laut galt.

• Die Bezeichnung des Motors ist gemäß
damaliger Rolls-Royce Tradition von einer
Vogelart, dem Merlinfohnen, übernommen
und nicht, wie oft vermutet, von dem
Zauberer Merlin.

English Translation
(via Google Translate)

• The Rolls-Royce Merlin is a 12-cylinder
  aircraft engine from Rolls-Royce V-type,
  which served many important British and
  American aircraft designs of World War II as
  a drive. From 1941 the engine was built
  under license by the Packard Motor Car
  Company in the U.S. as a Packard V-1650th.

• After the war, several passenger and cargo
  aircraft have been equipped with this engine,
such as Avro Lancastrian, Avro Tudor Avro
  York and, later, the Canadair C-4 (converted
  Douglas C-54). The civilian use of the Merlin
  was, however, limited as it remains robust,
  however, was too loud.

• The name of the motor is taken under the
  then Rolls-Royce tradition of one species, the
  Merlin falcon, and not, as often assumed, by
  the wizard Merlin.
Rolls-Royce Merlin Engine
(from German Wikipedia)

• Nach dem Krieg wurden diverse Passagier- und Frachtflugzeuge mit diesem Motor ausgestattet, so z. B. Avro Lancastrian, Avro Tudor und Avro York, später noch einmal die Canadair C-4 (umgebaute Douglas C-54). Der zivile Einsatz des Merlin hielt sich jedoch in Grenzen, da er als robust, aber zu laut galt.

English Translation (via Google Translate)

• The Rolls-Royce Merlin is a 12-cylinder aircraft engine from Rolls-Royce V-type, which served many important British and American aircraft designs of World War II as a drive. From 1941 the engine was built under license by the Packard Motor Car Company in the U.S. as a Packard V-1650th.
• After the war, several passenger and cargo aircraft have been equipped with this engine, such as Avro Lancastrian, Avro Tudor Avro York and, later, the Canadair C-4 (converted Douglas C-54). The civilian use of the Merlin was, however, limited as it remains robust, however, was too loud.
• The name of the motor is taken under the then Rolls-Royce tradition of one species, the Merlin falcon, and not, as often assumed, by the wizard Merlin.
Rolls-Royce Merlin Engine
(from German Wikipedia)


Turkish Translation
(via Google Translate)

- Rolls-Royce Merlin 12-den silindirli Rolls-Royce uçak motoru V tipi, bir sürücü olarak Dünya Savaşı’nın birçok önemli İngiliz ve Amerikan uçak tasarımları devam eder. 1.941 motor lisansı altında Packard Motor Car Company tarafından ABD’den Packard V olarak yapılmıştır Gönderen-1650
- Savaştan sonra, birkaç yolcu ve kargo uçakları ile Avro Lancastrian, Avro Avro York ve Tudor gibi bu motor, daha sonra, Canadair C-4 (Douglas C-54) dönüştürülmüş donatılmiştir. Olarak, ancak, çok yüksek olduğu sağlam kalır Merlin sivil kullanıma Ancak sınırlıydı.
- Motor adı daha sonra Rolls altında bir türün, Merlin şahin, ve değil-Royce geleneği, sıklıkta kabul, Merlin sihirbaz tarafından alınır.


Rolls-Royce Merlin Engine (from German Wikipedia)


• Nach dem Krieg wurden diverse Passagier- und Frachtflugzeuge mit diesem Motor ausgestattet, so z. B. Avro Lancastrian, Avro Tudor und Avro York, später noch einmal die Canadair C-4 (umgebaute Douglas C-54). Der zivile Einsatz des Merlin hielt sich jedoch in Grenzen, da er als robust, aber zu laut galt.


Arabic Translation (via Google Translate – 2017)
Machine Translation

• (Real-time speech-to-speech) Translation is a very demanding task
  – Simultaneous translators (in UN or EU Parliament) last about 30 minutes
  – Time pressure
  – Divergences between languages
    • German: Subject ........................... Verb
    • English:    Subject  Verb ............................
    • Arabic:     Verb Subject ..........................
Brief History

• 1950’s: Intensive research activity in MT
  – Translate Russian into English
• 1960’s: Direct word-for-word replacement
• 1966 (ALPAC): NRC Report on MT
  – Conclusion: MT no longer worthy of serious scientific investigation.
• 1966-1975: `Recovery period’
• 1975-1985: Resurgence (Europe, Japan)
• 1992-present: Resurgence (US)
  – Mostly Statistical Machine Translation since 1990s
  – Recently Neural Network/Deep Learning based machine translation
Early Rule-based Approaches

- Expert system-like rewrite systems
- Interlingua methods (analyze and generate)
- Information used for translation are compiled by humans
  - Dictionaries
  - Rules
Vauquois Triangle
Statistical Approaches

• Word-to-word translation
• Phrase-based translation
• Syntax-based translation (tree-to-tree, tree-to-string)
• Trained on parallel corpora
• Mostly noisy-channel (at least in spirit)
Deep Learning Approaches

• Models as a sequence to sequence mapping
• Recurrent networks
  – GRU/bi-LSTM
• Input represented with word/subword embeddings
• Output is decoded with Deep LMs, softmax/beam search
Early Hints on the Noisy Channel
Intuition

• “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’”

Warren Weaver

• (1955:18, quoting a letter he wrote in 1947)
Divergences between Languages

• Languages differ along many dimensions
  – Concept – Lexicon alignment – Lexical Divergence
  – Syntax – Structure Divergence
    • Word-order differences
      – English is Subject-Verb-Object
      – Arabic is Verb-Subject-Object
      – Turkish is Subject-Object-Verb
    • Phrase order differences
    • Structure-Semantics Divergences
Lexical Divergences

• English: wall
  – German: *Wand* for walls inside, *Mauer* for walls outside

• English: runway
  – Dutch: *Landingbaan* for when you are landing; *startbaan* for when you are taking off

• English: aunt
  – Turkish: *hala* (father’s sister), *teyze* (mother’s sister)

• Turkish: o
  – English: she, he, it
Lexical Divergences
How conceptual space is cut up
Lexical Gaps

• One language may not have a word for a concept in another language
  – Japanese: oyakoko
    • Best English approximation: “filial piety”
  – Turkish: gurbet
    • Where you are when you are not “home”
  – English: condiments
    • Turkish: ??? (things like mustard, mayo and ketchup)
Local Phrasal Structure Divergences

• English: a blue house
  – French: une maison bleu

• German: die ins Haus gehende Frau
  – English: the lady walking into the house
Structural Divergences

• English: I have a book.
  – Turkish: Benim kitabım var. (Lit: My book exists)

• French: Je m’appelle Jean (Lit: I call myself Jean)
  – English: My name is Jean.

• English: I like swimming.
  – German: Ich schwimme gerne. (Lit: I swim “likingly”.)
Major Rule-based MT Systems/Projects

• **Systran**
  – Major human effort to construct large translation dictionaires + limited word-reordering rules

• **Eurotra**
  – Major EU-funded project (1970s-1994) to translate among (then) 12 EC languages.
    • Bold technological framework
      – Structural Interlingua
    • Management failure
    • Never delivered a working MT system
    • Helped create critical mass of researchers
Major Rule-based MT Systems/Projects

- **METEO**
  - Successful system for French-English translation of Canadian weather reports (1975-1977)

- **PANGLOSS**
  - Large-scale MT project by CMU/USC-ISI/NMSU
  - Interlingua-based Japanese-Spanish-English translation
  - Manually developed semantic lexicons
Rule-based MT

• Manually develop rules to analyze the source language sentence (e.g., a parser)
  – => some source structure representation
• Map source structure to a target structure
• Generate target sentence from the transferred structure
Rule-based MT

Source language analysis

Noun Phrase
Pronoun
Verb
Verb Phrase
Noun Phrase
Sentence

Verb
Adj.

Noun
Noun Phrase

Verb Phrase
Sentence
Noun Phrase
Pronoun

Verb
Adj.

Noun
Noun Phrase

Verb
Adj.

Noun
Noun Phrase

Je
lire

livres
scientifiques

Syntactic Transfer
⇒

Swap

I read scientific books

lire livres scientifiques

Target language generation
Rules

• Rules to analyze the source sentences
  – (Usually) Context-free grammar rules coupled with linguistic features
    • Sentence => Subject-NP  Verb-Phrase
    • Verb-Phrase => Verb Object .....
Rules

• Lexical transfer rules
  – English: pound (N, monetary sense) => French: livre (N, feminine)
  – English: book (V) => French: réserver (V)

• Quite tricky for
Rules

• Structure Transfer Rules
  – English: $S \Rightarrow NP\ VP$
    French: $TR(S) \Rightarrow TR(NP)\ TR(VP)$
  – English: $NP \Rightarrow Adj\ Noun$
    French: $TR(NP) \Rightarrow Tr(Noun)\ Tr(Adj)$
    but there are exceptions for
    Adj=grand, petit, ....
Rules

Much more complex to deal with “real world” sentences.

Canadian Utilities had 1988 revenue of C$ 1.16 billion, mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers.
Example-based MT (EBMT)

• Characterized by its use of a bilingual corpus with parallel texts as its main knowledge base, at run-time.
• Essentially translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning.
• Find how (parts of) input are translated in the examples
  – Cut and paste to generate novel translations
Example-based MT (EBMT)

• Translation Memory
  – Store many translations,
    • source – target sentence pairs
  – For new sentences, find closes match
    • use edit distance, POS match, other similarity techniques
  – Do corrections,
    • map insertions, deletions, substitutions onto target sentence
  – Useful only when you expect same or similar sentence to show up again, but then high quality
Example-based MT (EBMT)

<table>
<thead>
<tr>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>• How much is that <strong>red umbrella</strong>?</td>
<td>• Ano <strong>akai kasa</strong> wa ikura desu ka?</td>
</tr>
<tr>
<td>• How much is that <strong>small camera</strong>?</td>
<td>• Ano <strong>chiisai kamera</strong> wa ikura desu ka?</td>
</tr>
<tr>
<td>• How much is that <strong>X</strong>?</td>
<td>• Ano <strong>X</strong> wa ikura desu ka?</td>
</tr>
</tbody>
</table>
Hybrid Machine Translation

• Use multiple techniques (rule-based/EBMT/Interlingua)

• Combine the outputs of different systems to improve final translations
How do we evaluate MT output?

- **Adequacy**: Is the meaning of the source sentence conveyed by the target sentence?
- **Fluency**: Is the sentence grammatical in the target language?
- These are rated on a scale of 1 to 5
How do we evaluate MT output?

<table>
<thead>
<tr>
<th></th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Je suis fatigué.</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Tired is I.</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Cookies taste good!</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>I am tired.</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
How do we evaluate MT output?

• This in general is very labor intensive
  – Read each source sentence
  – Evaluate target sentence for adequacy and fluency
• Not easy to do if you improve your MT system 100 times a day, and need to evaluate!
  – Could this be mechanized?
    • Later
MT Strategies (1954-2004)

Knowledge Acquisition Strategy

All manual

Electronic dictionaries

Hand-built by experts

Hand-built by non-experts

Original direct approach

Typical transfer system

Classic interlingual system

Shallow/ Simple

Word-based

Phrase tables

Learn from annotated data

Example-based MT

Learn from un-annotated data

Fully automated

Statistical MT

Syntactic Constituent Structure

Semantic analysis

Interlingua

Deep/ Complex

Knowledge Representation Strategy

New Research Goes Here!

Typical transfer system

Electronic dictionaries

Knowledge Acquisition Strategy

All manual
Statistical Machine Translation

• How does statistics and probabilities come into play?
  – Often statistical and rule-based MT are seen as alternatives, even opposing approaches – wrong !!!

<table>
<thead>
<tr>
<th></th>
<th>No Probabilities</th>
<th>Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Structure</td>
<td>EBMT</td>
<td>SMT</td>
</tr>
<tr>
<td>Deep Structure</td>
<td>Transfer Interlingua</td>
<td>Holy Grail</td>
</tr>
</tbody>
</table>

– Goal: structurally rich probabilistic models
Rule-based MT vs SMT

Expert System

Manually coded rules
If « ... » then ... 
If « ... » then ... 
...... 
Else ....

Statistical System

Bilingual parallel corpus

Statistical rules
P(but | mais)=0.7
P(however | mais)=0.3
P(where | où)=1.0
......

SignalWare: Mais où sont les neiges d’antan?

Expert system output
T: But where are the snows of yesteryear? P=0.41
T2: However, where are yesterday’s snows? P=0.33
T3: Hey - where did the old snow go? P=0.18
Data-Driven Machine Translation

Hmm, every time he sees “banco”, he either types “bank” or “bench” … but if he sees “banco de…”, he always types “bank”, never “bench”...

Man, this is so boring.

Translated documents

Slide by Kevin Knight
Statistical Machine Translation

• The idea is to use lots of parallel texts to model how translations are done.
  – Observe how words or groups of words are translated
  – Observe how translated words are moved around to make fluent sentences in the target sentences
1a. Garcia and associates.
1b. Garcia y asociados.

2a. Carlos Garcia has three associates.
2b. Carlos Garcia tiene tres asociados.

3a. his associates are not strong.
3b. sus asociados no son fuertes.

4a. Garcia has a company also.
4b. Garcia tambien tiene una empresa.

5a. its clients are angry.
5b. sus clientes estan enfadados.

6a. the associates are also angry.
6b. los asociados tambien estan enfadados.

7a. the clients and the associates are enemies.
7b. los clientes y los asociados son enemigos.

8a. the company has three groups.
8b. la empresa tiene tres grupos.

9a. its groups are in Europe.
9b. sus grupos estan en Europa.

10a. the modern groups sell strong pharmaceuticals.
10b. los grupos modernos venden medicinas fuertes.

11a. the groups do not sell zenzanine.
11b. los grupos no venden zanzanina.

12a. the small groups are not modern.
12b. los grupos pequenos no son modernos.
Parallel Texts

Clients do not sell pharmaceuticals in Europe

1a. Garcia and associates .
1b. Garcia y asociados .

2a. Carlos Garcia has three associates .
2b. Carlos Garcia tiene tres asociados .

3a. his associates are not strong .
3b. sus asociados no son fuertes .

4a. Garcia has a company also .
4b. Garcia tambien tiene una empresa .

5a. its clients are angry .
5b. sus clientes estan enfadados .

6a. the associates are also angry .
6b. los asociados tambien estan enfadados .

7a. the clients and the associates are enemies .
7b. los clientes y los asociados son enemigos .

8a. the company has three groups .
8b. la empresa tiene tres grupos .

9a. its groups are in Europe .
9b. sus grupos estan en Europa .

10a. the modern groups sell strong pharmaceuticals .
10b. los grupos modernos venden medicinas fuertes .

11a. the groups do not sell zenzanine .
11b. los grupos no venden zanzanina .

12a. the small groups are not modern .
12b. los grupos pequenos no son modernos .

Clientes no venden medicinas en Europa
Parallel Texts

1. employment rates are very low, especially for women.
2. the overall employment rate in 2001 was 46.8%.
3. the system covers insured employees who lose their jobs.
4. the resulting loss of income is covered in proportion to the premiums paid.
5. there has been no development in the field of disabled people.
6. overall assessment
7. no social dialogue exists in most private enterprises.
8. it should be reviewed together with all the social partners.
9. much remains to be done in the field of social protection.

1. istihdam oranları, özellikle kadınlar için çok düşüktür.
2. 2001 yılında genel istihdam oranı % 46,8' dir.
3. sistem, işini kaybeden sigortalı işsizleri kapsamaktadır.
4. ortaya çıkan gelir kaybı, ödenmiş primlerle orantılı olarak karşılanmaktadır.
5. engelli kişiler konusunda bir gelişme kaydedilmemiştir.
6. genel değerlendirme
7. özel işletmelerin çoğunuda sosyal diyalog yoktur.
8. konseyin yapısı, sosyal taraflar ile birlikte yeniden gözden geçirilmelidir.
9. sosyal koruma alanında yapılması gereken çok şey vardır.
Available Parallel Data (2004)

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

+ 1m-20m words for many language pairs

Millions of words (English side)
Available Parallel Data (2008)

• **Europarl**: 30 million words in 11 languages
• **Acquis Communitaire**: 8-50 million words in 20 EU languages
• **Canadian Hansards**: 20 million words from Canadian Parliamentary Proceedings
• **Chinese/Arabic - English**: over 100 million words from LDC
• Lots more French/English, Spanish/French/English from LDC
• Smaller corpora for many other language pairs
  – Usually English – Some other language.
Available Parallel Data (2017)

- The OPUS collection is growing! Check this page from time to time to see new data arriving...
- Contributions are very welcome! Please contact jorg.tiedemann@lingfil.uu.se.

Search & Browse
- OPUS multilingual search interface
- Europarl v7 search interface
- OpenSubtitles search interface
- ECB (European Central Bank)

Tools & Info
- OPUS Wiki
- Tools for tagging and parsing
- Downloads (tools and models)
- Other annotation and corpus tools
- Experimental visualization tool for monolingual and parallel text banks
- Upling at bitbucket
- A reliable Language Identification

Some Projects using OPUS
- LetSMT: On-line SMT toolkit
- CASMACAT: Computer-Aided Translation
- WMT: A conference on statistical MT
- Reverso: Translations in context
- SketchEngine: Tools for lexicographers
- sub-s-a-s: Translations in colloquial language

Links to other Resources
- The EuroParl corpus and WMT data
- CoDeEP: A cleaner and structured version of the Europarl corpus

Sub-corpora (downloads & info):
- Books: A collection of translated literature
- DOCC: Documents from the Catalan Government
- ECB: European Central Bank documents
- EMED: European Medicines Agency documents
- The EU bookshop corpus
- EUCorpus: The European constitution
- EuroParl: European Parliament Proceedings
- Gnome: GNOME localization files
- Global Voices: News stories in various languages
- DSL: Dutch Sampled corpus
- The Croatian - English WaC corpus
- IRC-Acquis: legislative EU texts
- KDE4: KDE4 localization files (v.2)
- KDE5: KDE5 manual corpus
- MBIS: Belgian Standardized corpus
- MultILU: Translated UN documents
- News Commentary (News Commentary 1999)
- OpenSubtitles (OpenSubtitles 2016)
- OpenSubtitles 2016: subtitles.org corpus
- OpenSubtitles 2016: subtitles.org 2016
- OpenSubtitles 2017: subtitles.org 2017
- OpenSubtitles 2018: subtitles.org 2018

n-20m words for many language pairs
Available Parallel Text

• A book has a few 100,000s words
• An educated person may read 10,000 words a day
  – 3.5 million words a year
  – 300 million words a lifetime
• Soon computers will have access to more translated text than humans read in a lifetime
More data is better!

- Language Weaver Arabic to English Translation

Description of the Iraqi President George Bush American elections—which will follow in the current month of the thirty—that they constitute a historic moment, recognizing that the organization of elections in current circumstances difficult issue.

It was considered bush in the press that the pronouncements of the possible organization of elections in most regions of the Iraqi punctually wish that the turnout where high. He added that "Iraqi 14 appear in the relative calm 18-II governorates

A description of the American president George W. Bush elections—Iraq, which will take place on the thirtieth session of the month-- as a historic moment, acknowledging that the organization of elections in the current difficult circumstances.

Bush said in press statements that it is possible to organize elections in most regions of Iraq to the deadline and I wish that the turnout are high. He added that "14 governorates of Iraq's 18 appeared in relative calm".

US President George W. Bush described Iraq elections—which will take place on the 30th of this month-- as a historic moment, acknowledging that the elections in the current situation is difficult.

Bush said in a press statement that it be possible to organize elections in most regions of Iraq in time and hoped that the rate of participation in the high. He added that "Iraqi 14 of the provinces of 18 appears to be relatively calm."

v.2.0 – October 2003

v.2.4 – October 2004

v.3.0 - February 2005
Sample Learning Curves

![Learning Curves Graph]

- Swedish/English
- French/English
- German/English
- Finnish/English

# of sentence pairs used in training

BLEU score

Experiments by Philipp Koehn
Preparing Data

• Sentence Alignment
• Tokenization/Segmentation
The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.
Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.
Sentence Alignment

- **1-1 Alignment**
  - 1 sentence in one side aligns to 1 sentence in the other side

- **0-n, n-0 Alignment**
  - A sentence in one side aligns to no sentences on the other side

- **n-m Alignment** *(n,m>0 but typically very small)*
  - n sentences on one side align to m sentences on the other side
Sentence Alignment

• Sentence alignments are typically done by dynamic programming algorithms
  – Almost always, the alignments are monotonic.
  – The lengths of sentences and their translations (mostly) correlate.
  – Tokens like numbers, dates, proper names, cognates help anchor sentences.
1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.
1. The old man is happy. He has fished many times.
2. His wife talks to him.
3. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Unaligned sentences are thrown out, and sentences are merged in n-to-m alignments (n, m > 0).
Tokenization (or Segmentation)

- English
  - Input (some byte stream):
    "There," said Bob.
  - Output (7 "tokens" or "words"):
    "There, " said Bob.

- Chinese
  - Input (byte stream):
    美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
  - Output:
    美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
The Basic Formulation of SMT

• Given a source language sentence $s$, what is the target language text $t$, that maximizes

$$p(t \mid s)$$

• So, any target language sentence $t$ is a “potential” translation of the source sentence $s$
  – But probabilities differ
  – We need that $t$ with the highest probability of being a translation.
The Basic Formulation of SMT

- Given a source language sentence $s$, what is the target language text $t$, that maximizes

  $$p(t \mid s)$$

- We denote this computation as a search

  $$t^* = \text{argmax}_t p(t \mid s)$$
The Basic Formulation of SMT

- We need to compute \( t^* = \text{argmax}_t p(t \mid s) \)

- Using Bayes’ Rule we can “factorize” this into two separate problems

\[
\begin{align*}
t^* &= \text{argmax}_t \frac{p(s \mid t)p(t)}{p(s)} \\
&= \text{argmax}_t p(s \mid t)p(t)
\end{align*}
\]

- Search over all possible target sentences \( t \)
  - For a given \( s, p(s) \) is constant, so no need to consider it in the maximization
The Noisy Channel Model

(Target) Dün Ali’yi gördüm.

(Source) I saw Ali yesterday

What was target sentence he used?

What are likely sentences he could have said in the target language?

How could the channel have “corrupted” target to source language?
Dün Ali’yi gördüm.

I saw Ali yesterday.

Neleri söylemesi olası?

Kanal söyleneni ne şekilde “bozmuş” olabilir?

Acaba ne söyledi?
Where do the probabilities come from?

Source(s)/Target(t)
Bilingual Text

Target
Text

Statistical Analysis

Source

Broken
Target

Translation Model $P(S|T)$

Source Sentence

Decoding algorithm
$\arg\max_T P(T) \times P(S|T)$

Target

Language Model $P(T)$

Target Sentence
The Statistical Models

• **Translation model** $p(S|T)$
  – Essentially models **Adequacy** without having to worry about **Fluency**.
    • $p(S|T)$ is high for sentences $S$, if words in $S$ are in general translations of words in $T$.

• **Target Language Model** $p(T)$
  – Essentially models **Fluency** without having to worry about **Adequacy**
    • $p(T)$ is high if a sentence $T$ is a fluent sentence in the target language
How do the models interact?

• Maximizing $p(S \mid T) P(T)$
  – $p(T)$ models “good” target sentences (Target Language Model)
  – $p(S\mid T)$ models whether words in source sentence are “good” translation of words in the target sentence (Translation Model)

<table>
<thead>
<tr>
<th>I saw Ali yesterday</th>
<th>Good Target? $P(T)$</th>
<th>Good match to Source? $P(S\mid T)$</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bugün Ali’ye gittim</td>
<td>✔️</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Okulda kalmışlar</td>
<td>✔️</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Var gelmek ben</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
<tr>
<td>Dün Ali’yi görüdüm</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Gördüm ben dün Ali’yi</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Dün Ali’ye görüdüm</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
</tr>
</tbody>
</table>
Three Problems for Statistical MT

• **Language model**
  – Given a target sentence $T$, assigns $p(T)$
    * good target sentence -> high $p(T)$
    * word salad -> low $p(T)$

• **Translation model**
  – Given a pair of strings $<S,T>$, assigns $p(S \mid T)$
    * $<S,T>$ look like translations -> high $p(S \mid T)$
    * $<S,T>$ don’t look like translations -> low $p(S \mid T)$

• **Decoding algorithm**
  – Given a language model, a translation model, and a new sentence $S$ ... find translation $T$ maximizing $p(T) \times p(S \mid T)$
The Classic Language Model: Word n-grams

• Helps us choose among sentences
  – He is on the soccer field
  – He is in the soccer field

  – Is table the on cup the
  – The cup is on the table

  – Rice shrine
  – American shrine
  – Rice company
  – American company
The Classic Language Model

• What is a “good” target sentence? (HLT Workshop 3)
• \( T = t_1 \ t_2 \ t_3 \ ... \ t_n; \)
• We want \( P(T) \) to be “high”
• A good approximation is by short n-grams
  \[
  P(T) \approx P(t_1|\text{START}) \cdot P(t_2|\text{START},t_1) \cdot P(t_3|t_1,t_2) \cdot \ldots \cdot P(t_i|t_{i-2},t_{i-1}) \cdot \\
  \ldots \cdot P(t_n|t_{n-2},t_{n-1})
  \]

  – Estimate from large amounts of text
    • Maximum-likelihood estimation
    • Smoothing for unseen data
      – You can never see all of language
    • There is no data like more data (e.g., \( 10^9 \) words would be nice)
The Classic Language Model

• If the target language is English, using 2-grams

\[ P(I \text{ saw water on the table}) \approx P(I | \text{START}) \times P(\text{saw} | I) \times P(\text{water} | \text{saw}) \times P(\text{on} | \text{water}) \times P(\text{the} | \text{on}) \times P(\text{table} | \text{the}) \times P(\text{END} | \text{table}) \]
The Classic Language Model

- If the target language is English, using 3-grams

\[
P(I \text{ saw water on the table}) \approx
\]
\[
P(I \mid \text{START, START}) \times
P(\text{saw} \mid \text{START, I}) \times
P(\text{water} \mid \text{I, saw}) \times
P(\text{on} \mid \text{saw, water}) \times
P(\text{the} \mid \text{water, on}) \times
P(\text{table} \mid \text{on, the}) \times
P(\text{END} \mid \text{the, table})
\]
Translation Model?

Generative approach:

Mary did not slap the green witch

Source-language morphological structure

Source parse tree

Semantic representation

Target structure

Maria no dio una botefada a la bruja verde

What are all the possible moves and their associated probability tables?
The Classic Translation Model
Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative approach:

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 dió una botefada a la verde bruja

Maria no dió una botefada a la bruja verde

Predict count of target words
Predict target words from NULL
Translate source to target words
Reorder target words
The Classic Translation Model
Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative approach:

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 dió una botefada a la verde bruja

Maria no dió una botefada a la bruja verde

Selected as the most likely by P(T)
Basic Translation Model (IBM M-1)

- Model $p(t \mid s, m)$
  - $t = <t_1, t_2, ..., t_m>$, $s = <s_1, s_2, ..., s_n>$
- Lexical translation makes the following assumptions
  - Each word $t_i$ in $t$ is generated from exactly one word in $s$.
  - Thus, we have a latent alignment $a_i$ that indicates which word $t_i$ “came from.” Specifically it came from $t_{a_i}$.
  - Given the alignments $a$, translation decisions are conditionally independent of each other and depend only on the aligned source word $t$. 

Basic Translation Model (IBM M-1)

\[ p(t|s, m) = \sum_{a \in [0,n]^m} p(a | s, m) \times \prod_{i=1}^{m} p(t_i | s_{a_i}) \]

- \( p(\text{alignment}) \)
- \( p(\text{translation | alignment}) \)
Parameters of the IBM 3 Model

• **Fertility**: How many words does a source word get translated to?
  
  – \( n(k \mid s) \): the probability that the source word \( s \) gets translated as \( k \) target words
  
  – Fertility depends solely on the source words in question and not other source words in the sentence, or their fertilities.

• **Null Probability**: What is the probability of a word magically appearing in the target at some position, without being the translation of any source word?
  
  – P-null
Parameters of the IBM 3 Model

• **Translation**: How do source words translate?
  – \( \text{tr}(t | s) \): the probability that the source word \( s \) gets translated as the target word \( t \)
  – Once we fix \( n(k | s) \) we generate \( k \) target words

• **Reordering**: How do words move around in the target sentence?
  – \( d(j | i) \): distortion probability – the probability of word at position \( i \) in a source sentence being translated as the word at position \( j \) in target sentence.
    • Very dubious!!
How IBM Model 3 works

1. For each source word $s_i$ indexed by $i = 1, 2, ..., m$, choose fertility $\phi_i$ with probability $n(\phi_i | s_i)$.

2. Choose the number $\phi_0$ of “spurious” target words to be generated from $s_0 = \text{NULL}$
How IBM Model 3 works

3. Let $q$ be the sum of fertilities for all words, including NULL.

4. For each $i = 0, 1, 2, \ldots, m$, and each $k = 1, 2, \ldots, \phi_i$, choose a target word $t_{ik}$ with probability $tr(t_{ik} \mid s_i)$.

5. For each $i = 1, 2, \ldots, l$, and each $k = 1, 2, \ldots, \phi_i$, choose target position $p_{ik}$ with probability $d(p_{ik} \mid i, l, m)$. 
How IBM Model 3 works

6. For each $k = 1, 2, \ldots, \phi_0$, choose a position $\pi_{0k}$ from the remaining vacant positions in 1, 2, ... $q$, for a total probability of $1/\phi_0$.

7. Output the target sentence with words $t_{ik}$ in positions $\pi_{ik}$ (0 <= $i$ <= $m$, 1 <= $k$ <= $\phi_i$).
Example

- $n$-parameters
- $n(0,b)=0$, $n(1,b)=2/2=1$
- $n(0,c)=1/1=1$, $n(1,c)=0$
- $n(0,d)=0$, $n(1,d)=1/2=0.5$, $n(2,d)=1/2=0.5$
### Example

<table>
<thead>
<tr>
<th>b</th>
<th>c</th>
<th>d</th>
<th>b</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- t-parameters
- $t(x \mid b)=1.0$
- $t(y \mid d)=2/3$
- $t(z \mid d)=1/3$
### Example

<table>
<thead>
<tr>
<th>b</th>
<th>c</th>
<th>d</th>
<th>b</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>++</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>y</td>
<td>z</td>
<td>x</td>
<td>y</td>
</tr>
</tbody>
</table>

- **d-parameters**
  - \(d(1|1,3,3)=1.0\)
  - \(d(1|1,2,2)=1.0\)
  - \(d(2|2,3,3)=0.0\)
  - \(d(3|3,3,3)=1.0\)
  - \(d(2|2,2,2)=1.0\)
Example

- \( p_1 \)
- No target words are generated by NULL so \( p_1 = 0.0 \)
The Classic Translation Model
Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative approach:

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 dió una botefada a la verde bruja

Selected as the most likely by P(T)
How do we get these parameters?

• Remember we had aligned parallel sentences

• Now we need to figure out how words align with other words.
  – Word alignment
Word Alignments

- One source word can map to 0 or more target words
  - But not vice versa
    - technical reasons
- Some words in the target can magically be generated from an invisible NULL word
- A target word can only be generated from one source word
  - technical reasons
Word Alignments

\[ tr(\text{oeuvre} \mid \text{worked}) = \frac{c(\text{oeuvre} \mid \text{worked})}{c(\text{worked})} \]

- Count over all aligned sentences
- worked
  - fonctionné(30), travaillé(20), marché(27), oeuvré (13)
  - \( tr(\text{oeuvre} \mid \text{worked}) = 0.13 \)
- Similarly, \( n(3, \text{many}) \) can be computed.
How do we get these alignments?

• We only have aligned sentences and the constraints:
  – One source word can map to 0 or more target words
    • But not vice versa
  – Some words in the target can magically be generated from an invisible NULL word
  – A target word can only be generated from one source word
• Estimation – Maximization Algorithm
  – Mathematics is rather complicated
How do we get these alignments?

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

All word alignments equally likely

All $p($french-word $| $english-word$)$ equally likely
How do we get these alignments?

... la maison ... la maison bleue ... la fleur ...

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

“la” and “the” observed to co-occur frequently, so $p(\text{la} \mid \text{the})$ is increased.
How do we get these alignments?

... la maison ... la maison bleue ... la fleur ...

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

“house” co-occurs with both “la” and “maison”, but $p(\text{maison} \mid \text{house})$ can be raised without limit, to 1.0, while $p(\text{la} \mid \text{house})$ is limited because of “the”

(pigeonhole principle)
How do we get these alignments?

... la maison ... la maison bleue ... la fleur ... 

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

settling down after another iteration
How do we get these alignments?

... la maison ... la maison bleue ... la fleur ...

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

Inherent hidden structure revealed by EM training!
For further details, see:

• “A Statistical MT Tutorial Workbook” (Knight, 1999).
• “The Mathematics of Statistical Machine Translation” (Brown et al, 1993)
• Software: GIZA++
Decoding for “Classic” Models

• Of all conceivable English word strings, find the one maximizing $p(t) \times p(t \mid s)$

• Decoding is an NP-complete challenge
  – Reduction to Traveling Salesman problem (Knight, 1999)

• Several search strategies are available

• Each potential target output is called a *hypothesis*.
Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001)]
Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

The Classic Results

- *la politique de la haine.* (Original Source)
- politics of hate. (Reference Translation)
- the policy of the hatred. (IBM4+N-grams+Stack)

- *nous avons signé le protocole.* (Original Source)
- we did sign the memorandum of agreement. (Reference Translation)
- we have signed the protocol. (IBM4+N-grams+Stack)

- *où était le plan solide ?* (Original Source)
- but where was the solid plan? (Reference Translation)
- where was the economic base? (IBM4+N-grams+Stack)

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and
Flaws of Word-Based MT

- Multiple source words for one target word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - “real estate”, “note that”, “interest in”
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place
Phrase-Based Statistical MT

- Source input segmented into phrases
  - “phrase” is any sequence of words
- Each phrase is probabilistically translated into target
  - $P(\text{to the conference} \mid \text{zur Konferenz})$
  - $P(\text{into the meeting} \mid \text{zur Konferenz})$
- Phrases are probabilistically re-ordered
Advantages of Phrase-Based SMT

• Many-to-many mappings can handle non-compositional phrases

• Local context is very useful for disambiguating
  – “Interest rate”  →  ...
  – “Interest in”  →  ...

• The more data, the longer the learned phrases
  – Sometimes whole sentences
How to Learn the Phrase Translation Table?

• One method: “alignment templates”
• Start with word alignment, build phrases from that.

Mary did not slap the green witch

This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.

This is the best (or “Viterbi”) alignment.
How to Learn the Phrase Translation Table?

• One method: “alignment templates” (Och et al, 1999)
• Start with word alignment, build phrases from that.
IBM Models are 1-to-Many

• Run IBM-style aligner both directions, then merge:

T → S best alignment

S → T best alignment

MERGE

Union or Intersection
How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*
Phrase alignment must contain all alignment points for all the words in both phrases!
Maria did not slap the green witch.

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
Maria did not slap the green witch.

Word Alignment Induced Phrases:

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(a la, the) (dió una bofetada a, slap the)
Mary did not slap the green witch.

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)
 bruja verde, green witch)
Mary did not slap the green witch.
Mary did not slap the green witch.

Word Alignment Induced Phrases:

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)
(a la, the) (dió una bofetada a, slap the)
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)
(bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap)
(a la bruja verde, the green witch) (Maria no dió una bofetada a la, Mary did not slap the)
(no dió una bofetada a la, did not slap the) (dió una bofetada a la bruja verde, slap the green witch)
Mary did not slap the green witch.

(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green) (a la, the) (dió una bofetada a, slap the) (Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the) (bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap) (a la bruja verde, the green witch) (Maria no dió una bofetada a la, Mary did not slap the) (no dió una bofetada a la, did not slap the) (dió una bofetada a la bruja verde, slap the green witch) (Maria no dió una bofetada a la bruja verde, Mary did not slap the green witch)
Phrase Pair Probabilities

• A certain phrase pair (s-s-s, t-t-t) may appear many times across the bilingual corpus.
  - We hope so!

• So, now we have a vast list of phrase pairs and their frequencies – how to assign probabilities?
Phrase-based SMT

• After doing this to millions of sentences
  – For each phrase pair \((t, s)\)
    • Count how many times \(s\) occurs
    • Count how many times \(s\) is translated to \(t\)
    • Estimate \(p(t \mid s)\)
Decoding

• During decoding
  – a sentence is segmented into “phrases” in all possible ways
  – each such phrase is then “translated” to the target phrases in all possible ways
  – Translations are also moved around
  – Resulting target sentences are scored with the target language model

• The decoder actually does NOT actually enumerate all possible translations or all possible target sentences
  – Pruning
Decoding
Basic Model, Revisited

\[ \text{argmax } P(t \mid s) = t \]

\[ \text{argmax } P(t) \times P(s \mid t) / P(s) = t \]

\[ \text{argmax } P(t) \times P(t \mid s) = t \]
Basic Model, Revisited

$$\text{argmax } P(t \mid s) = t$$

$$\text{argmax } P(t) \times P(s \mid t) / P(s) = t$$

$$\text{argmax } P(t)^{2.4} \times P(t \mid s) \text{ seems to work better}$$
argmax $P(t \mid s)$ =

$t$

argmax $P(t) \times P(s \mid t) / P(s)$ =

$t$

argmax $P(t)^{2.4} \times P(t \mid s) \times \text{length}(t)^{1.1}$

$t$

Rewards longer hypotheses, since these are unfairly punished by $p(t)$
Basic Model, Revisited

\[
\text{argmax } P(t)^{2.4} \times P(s \mid t) \times \text{length}(t)^{1.1} \times \text{KS}^{3.7} \ldots
\]

Lots of knowledge sources vote on any given hypothesis.

“Knowledge source” = “feature function” = “score component”.

Feature function simply scores a hypothesis with a real value.

(May be binary, as in “e has a verb”).

Problem: How to set the exponent weights?
Maximum BLEU Training

Learning Algorithm for Directly Reducing Translation Error Yields big improvements in quality.
Automatic Machine Translation Evaluation

- Objective
- Inspired by the Word Error Rate metric used by ASR research
- Measuring the “closeness” between the MT hypothesis and human reference translations
  - Precision: n-gram precision
  - Recall:
    - Against the best matched reference
    - Approximated by brevity penalty
- Cheap, fast
- Highly correlated with human evaluations
- MT research has greatly benefited from automatic evaluations
- Typical metrics: BLEU, NIST, F-Score, Meteor, TER
**BLEU Evaluation**

**Reference (human) translation:**
The US island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself Osama Bin Laden and threatening a biological/chemical attack against the airport.

**Machine translation:**
The American [?] International airport and its the office at [?] receives one calls self the sand Arab rich business [?] and so on electronic mail, which sends out; The threat will be able after the maintenance at the airport.

**N-gram precision (score between 0 & 1)**
- what % of machine n-grams (a sequence of words) can be found in the reference translation?

**Brevity Penalty**
- Can’t just type out single word “the” (precision 1.0!)

Extremely hard to trick the system, i.e. find a way to change MT output so that BLEU score increases, but quality doesn’t.
Reference translation 1:
The US island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself Osama Bin Laden and threatening a biological/chemical attack against the airport.

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the rich Saudi Arabian businessman Osama Bin Laden and that threatened to launch a biological and chemical attack on the airport.

Machine translation:
The American [?] International airport and its the office to [?] receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out; The threat will be able after the maintenance at the airport to start the biochemistry attack.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden, which threatens to launch a biochemical attack on airport. Guam authority has been on alert.

Reference translation 4:
US Guam International Airport and its offices received an email from Mr. Bin Laden and other rich businessmen from Saudi Arabia. They said there would be biochemistry air raid to Guam Airport. Guam needs to be in high precaution about this matter.
BLEU in Action

- Reference Translation: *The gunman was shot to death by the police.*

- The gunman was shot kill.
- Wounded police jaya of
- The gunman was shot dead by the police.
- The gunman arrested by police kill.
- The gunmen were killed.
- The gunman was shot to death by the police.
- The ringer is killed by the police.
- Police killed the gunman.

- Green = 4-gram match (good!)  Red = unmatched word (bad!)
BLEU Formulation

\[
BLEU = \min \left( 1, \frac{\text{output} - \text{length}}{\text{reference} - \text{length}} \right) \left( \prod_{i=1}^{4} \text{precision}_i \right)^{\frac{1}{4}}
\]

\text{precision}_i: i\text{-gram precision over the whole corpus}
Correlation with Human Judgment
What About Morphology?

• Issue for handling morphologically complex languages like Turkish, Hungarian, Finnish, Arabic, etc.
  – A word contains much more information than just the root word

  • **Arabic**: \(\text{wsyktbunha (wa+sa+ya+ktub+ūn+ha “and they will write her”)}\)
    – What are the alignments?

  • **Turkish**: \(\text{gelenbilecekmissin (gel+ebil+ecek+mis+sin (I heard) you would be coming))}\)
    – What are the alignments?
Morphology & SMT

- Finlandiyanılaştıramadıklarınızdanmişsınızcasına

- Finlandiya+lı+laş+tır+ama+dık+lar+ımız+dan+miş+sınızcasına

- (behaving) as if you have been one of those whom we could not convert into a Finn(ish citizen)/someone from Finland
Most of the time, the morpheme order is “reverse” of the corresponding English word order

- yapabileceksek
  - yap+abil+ecek+se+k
  - if we will be able to do (something)

- yaptıtabildiğimizde
  - yap+tir+t+tığ+ımız+da
  - when/at the time we had (someone) have (someone else) do (something)

- görüntülenebilir
  - görüntüle+n+ebil+ir
  - it can be visualize+d

- sakarlıklarından
  - sakar+lık+ları+ndan
  - of/from/due-to their clumsi+ness
Morphology and Alignment

• Remember the alignment needs to count co-occurring words
  – If one side of the parallel text has little morphology (e.g. English)
  – The other side has lots of morphology
• Lots of words on the English side either don’t align or align randomly
Morphology & SMT

- If we ignore morphology
  - Large vocabulary size on the Turkish side
  - Potentially noisy alignments
  - The link **activity**-**faaliyet** is very “loose”

<table>
<thead>
<tr>
<th>Word Form</th>
<th>Count</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>faaliyet</td>
<td>3</td>
<td><strong>activity</strong></td>
</tr>
<tr>
<td>faaliyete</td>
<td>1</td>
<td>to the <strong>activity</strong></td>
</tr>
<tr>
<td>faaliyetinde</td>
<td>1</td>
<td>in its <strong>activity</strong></td>
</tr>
<tr>
<td>faaliyetler</td>
<td>3</td>
<td><strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetlere</td>
<td>6</td>
<td>to the <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetleri</td>
<td>7</td>
<td>their <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetlerin</td>
<td>7</td>
<td>of the <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetlerinde</td>
<td>1</td>
<td>in their <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetlerine</td>
<td>5</td>
<td>to their <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetlerini</td>
<td>1</td>
<td>their <strong>activities</strong> (accusative)</td>
</tr>
<tr>
<td>faaliyetlerinin</td>
<td>2</td>
<td>of their <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyetleriyle</td>
<td>1</td>
<td>with their <strong>activities</strong></td>
</tr>
<tr>
<td>faaliyette</td>
<td>2</td>
<td>in (the) <strong>activity</strong></td>
</tr>
<tr>
<td>faaliyetteki</td>
<td>1</td>
<td>that is in <strong>activity</strong></td>
</tr>
</tbody>
</table>

**TOTAL** 41
An Example E – T Translation

we are going to your hotel in Taksim by taxi

we are going to your hotel in Taksim by taxi
we are going to your hotel in Taksim by taxi
we are going to your hotel in Taksim by taxi

we are going to your hotel in Taksim by taxi
we are going to your hotel in Taksim by taxi
we are going to your hotel in Taksim by taxi

we are going to your hotel in Taksim by taxi
Morphology and Parallel Texts

• Use
  – Morphological analyzers (HLT Workshop 2)
  – Tagger/Disambiguators (HLT Workshop 3)
• to split both sides of the parallel corpus into morphemes
Morphology and Parallel Texts

• A typical sentence pair in this corpus looks like the following:

  • Turkish:
    - kat +hl +ma ortaklık +sh +nhn uygula +hn +ma +sh , ortaklık anlaşma +sh çerçeve +sh +nda izle +hn +yacak +dhr .

  • English:
    - the implementation of the accession partnership will be monitor +ed in the framework of the association agreement
Results

• Using morphology in Phrase-based SMT certainly improves results compared to just using words
• But
  – Sentences get much longer and this hurts alignment
  – We now have an additional problem: getting the morpheme order on each word right
Syntax and Morphology Interaction

• A completely different approach
  – Instead of dividing up Turkish side into morpheme
  – Collect “stuff” on the English side to make-up “words”.
  – What is the motivation?
Suppose we can do some **syntactic analysis on the English side**.
we are going to your hotel in Taksim by taxi

• to your hotel
  – to is the preposition related to hotel
  – your is the possessor of hotel

• to your hotel => hotel +your+to
  otel +iniz+e
  – separate content from local syntax
Syntax and Morphology Interaction

we are go+ing to your hotel in Taksim by taxi

• we are go+ing
  – we is the subject of go
  – are is the auxiliary of go
  – ing is the present tense marker for go

• we are go+ing => go +ing+are+we
  gid +iyor+uz
  – separate content from local syntax
we are going to your hotel in Taksim by taxi

Now align only based on root words – the syntax alignments just follow that
Syntax and Morphology Interaction
Syntax and Morphology Interaction

• Transformations on the English side reduce sentence length
• This helps alignment
  – Morphemes and most function words never get involved in alignment
• We can use factored phrase-based translation
  – Phrased-based framework with morphology support
Syntax and Morphology Interaction

Experiments

Number of Tokens

BLEU Scores

English

Turkish

BLEU Score
Syntax and Morphology Interaction

• She is reading.
  – She is the subject of read
  – is is the auxiliary of read

• She is read+ing => read +ing+is+she
  taQrAA QrAA +*ta
Neural Machine Translation
Teşekkürler/Thanks
MT Strategies (1954-2004)

Knowledge Acquisition Strategy

- All manual
- Hand-built by experts
- Hand-built by non-experts
- Original direct approach
- Typical transfer system
- Classic interlingual system

Knowledge Representation Strategy

- Shallow/ Simple
- Word-based only
- Phrase table
- Learn from annotated data
- Deep/ Complex
- Interlingua
- Semantic analysis
- Syntactic Constituent Structure

Statistical MT
- Learn from un-annotated data

Example-based MT
- Phrase table
- Learn from annotated data

Electronic dictionaries

Typical transfer system

New Research Goes Here!

Slide by Laurie Gerber
Syntax in SMT

• Early approaches relied on high-performance parsers for one or both languages
  – Good applicability when English is the source language
    • Tree-to-tree or tree-to-string transductions

• Recent approaches induce synchronous grammars during training
  – Grammar that describe two languages at the same time
    • NP => ADJe1 NPee2 : NPe2 ADJe1
Kare ha ongaku wo kiku no ga daisuki desu
Tree-to-String Transformation

- Each step is described by a statistical model
  - Reorder children on a node probabilistically
  - R-table
  - English – Japanese table

| Original Order   | Reordering   | P(reorder|original) |
|------------------|--------------|--------------|
| PRP VB1 VB2      | PRP VB1 VB2  | 0.074        |
|                  | PRP VB2 VB1  | 0.723        |
|                  | VB1 PRP VB2  | 0.061        |
|                  | VB1 VB2 PRP  | 0.037        |
|                  | VB2 PRP VB1  | 0.083        |
|                  | VB2 VB1 PRP  | 0.021        |
| VB TO            | VB TO        | 0.107        |
|                  | TO VB        | 0.893        |
| TO NN            | TO NN        | 0.251        |
|                  | NN TO        | 0.749        |
Tree-to-String Transformation

• Each step is described by a statistical model
  – Insert new sibling to the left or right of a node probabilistically
  – Translate source nodes probabilistically
Hierarchical phrase models

• Combines phrase-based models and tree structures
• Extract synchronous grammars from parallel text
• Uses a statistical chart-parsing algorithm during decoding
  – Parse and generate concurrently
For more info

• Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009
  – http://aclweb.org/anthology-new/W/W09/#2300

• Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)
  – http://aclweb.org/anthology-new/W/W08/#0400
Acknowledgments

• Some of the tutorial material is based on slides by
  – Kevin Knight (USC/ISI)
  – Philipp Koehn (Edinburgh)
  – Reyyan Yeniterzi (CMU/LTI)
Important References

• **Statistical Machine Translation (2010)**
  – Philipp Koehn
  – Cambridge University Press

• **Neural Machine Translation (2018)**

• **SMT Workbook (1999)**
  – Kevin Knight
  – Unpublished manuscript at [http://www.isi.edu/~knight/](http://www.isi.edu/~knight/)

• [http://www.statmt.org](http://www.statmt.org)
  – Look for “Workshop on Statistical Machine Translation”