

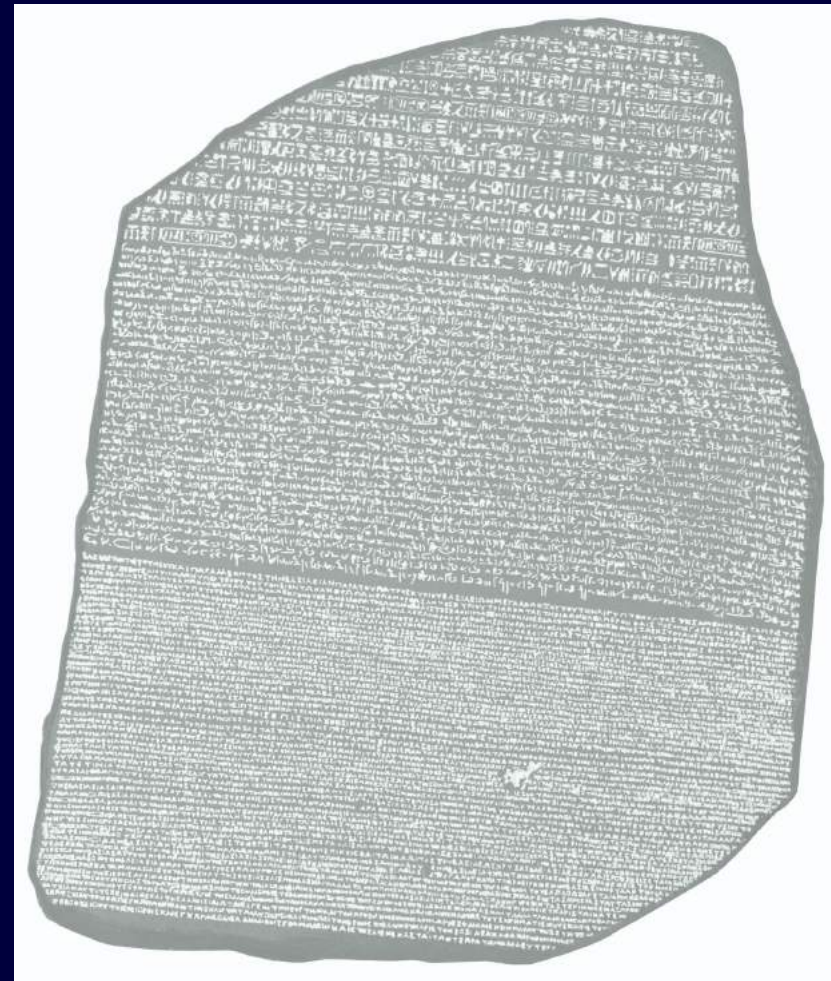
(Mostly Statistical) Machine Translation

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

Machine Translation

Many possible legitimate translations!

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israel government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport's security.

Israel is responsible for safety work at this airport.

Israel presides over the security of the airport.

Israel took charge of the airport security.

The safety of this airport is taken charge of by Israel.

This airport's security is the responsibility of the Israeli security officials.

Machine Translation

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 Merlinfalken, 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.

Machine Translation

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 Merlinfalken, 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.

Machine Translation

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 Merlinfalken, übernommen und nicht, wie oft vermutet, von dem Zauberer Merlin.

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'de Packard V olarak yaptırılmıştır 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) dönüştürülür donatılmıştır. Olarak, ancak, çok yüksek oldu 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.

Machine Translation

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 Lancaster, 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 Merlinfalken, übernommen und nicht, wie oft vermutet, von dem Zauberer Merlin.

Arabic Translation (via Google Translate -- 2009)

البريطانية والأمريكية تصممات الطائرات في م
من عام الحرب العالمية الثانية بمثابة محرك الأقراص.
1941 وكان المحرك الذي يتم بناؤه بموجب ترخيص من
قبل باكارد السيارات من شركة في الولايات المتحدة
1650 - باعتباره الخامس باكارد
بعد الحرب ، والعدد من الركاب والطائرات الشحن تم
تزويد هذا المحرك ، مثل أفرو Lancastrian ، أفرو
تيودور أفرو ونويويورك ، في وقت لاحق ، وجيم
استخدم (4 - 54) تحويل جيم دوغلاس - Canadair
المدني في ميرلين بيدي أنه كان محدودا كما هو عليه لا
نعالها جدا. يزال قويا ، ولكن ، كما
اسم المحرك هو المتخذة في إطار رولز رويس ثم تقلد
لأحد الأنواع ، الصقر ميرلين ، وليس ، كما يفترض
لج. افي كثير من الأحيان ، من قبل ميرلين المع

Machine Translation

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 Merlinfalken, übernommen und nicht, wie oft vermutet, von dem Zauberer Merlin.

Arabic Translation (via Google Translate – 2017)

و رولز رويس ميرلين هو محرك رولز رويس الخامس من نوع 12 اسطوانة التي تعمل على العديد من كبرى الحرب العالمية الثانية تصاميم الطائرات البريطانية والأمريكية. من عام 1941، تم بناء المحرك بموجب ترخيص من شركة باكارد للسيارات في الولايات المتحدة وباكارد V-1650.

بعد الحرب، وقد تم تجهيز مختلف طائرات الركاب والبضائع مع هذا المحرك، مثل: كما أفرو لانكاستريان، أفرو تودور و أفرو يورك، في وقت لاحق مرة أخرى ككنداير C-4 (تحويل دوغلاس C-54). ومع ذلك، كانت مهمة مدنية ميرلين محدودة، حيث اعتبر قوية، ولكن بصوت عال جدا.

اسم المحرك هو وفقا لتقاليد رولز رويس ثم من أنواع الطيور، والصقور ميرلين، واعتمدت وليس، كما يقتض في كثير من الأحيان من قبل ميرلين الساحر.

☆ □ ◀ ▶ ✎

w rulz ruis mirilin hu muhrak rulz ruis alkhamis min nawe 12 aslswanat
alty taamal aslaa aledyd min kubraa alharb alealamiat althdhanat
tasamim alttayirat albritaniat wal'amrikiati. min eam 1941, tama bina'
almaharik bmwbj tarkhis min sharikat biakard lilalyaat fi alwilaayat
almutahidat wabiakard V-1650.

baed alharb, waqad tama tajhiz mukhtalif tayirat alrukab walbadayie
mae hdha almuhraki, mithl. kama 'afro lankastarian, 'afro tudur w 'afro
yurk, fi waqt lahiq maratan 'ukhraa kanadir C-4 (thawil dwghias C-54).
wamae dhik, kanat muhimatan madaniatan muyriliin mahdudatan, hayth
aue tubir qawiat, walakun bisawt eal jiddaan.

aism almuharik hu wifqaan litaqalid rulz rawis thuma min 'anwae
alttayuri, walsuqur mirilin, waietamadat walaaysa, kama yufarad fi kthyr
min afahyan min qibal mirin alsaahir.

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

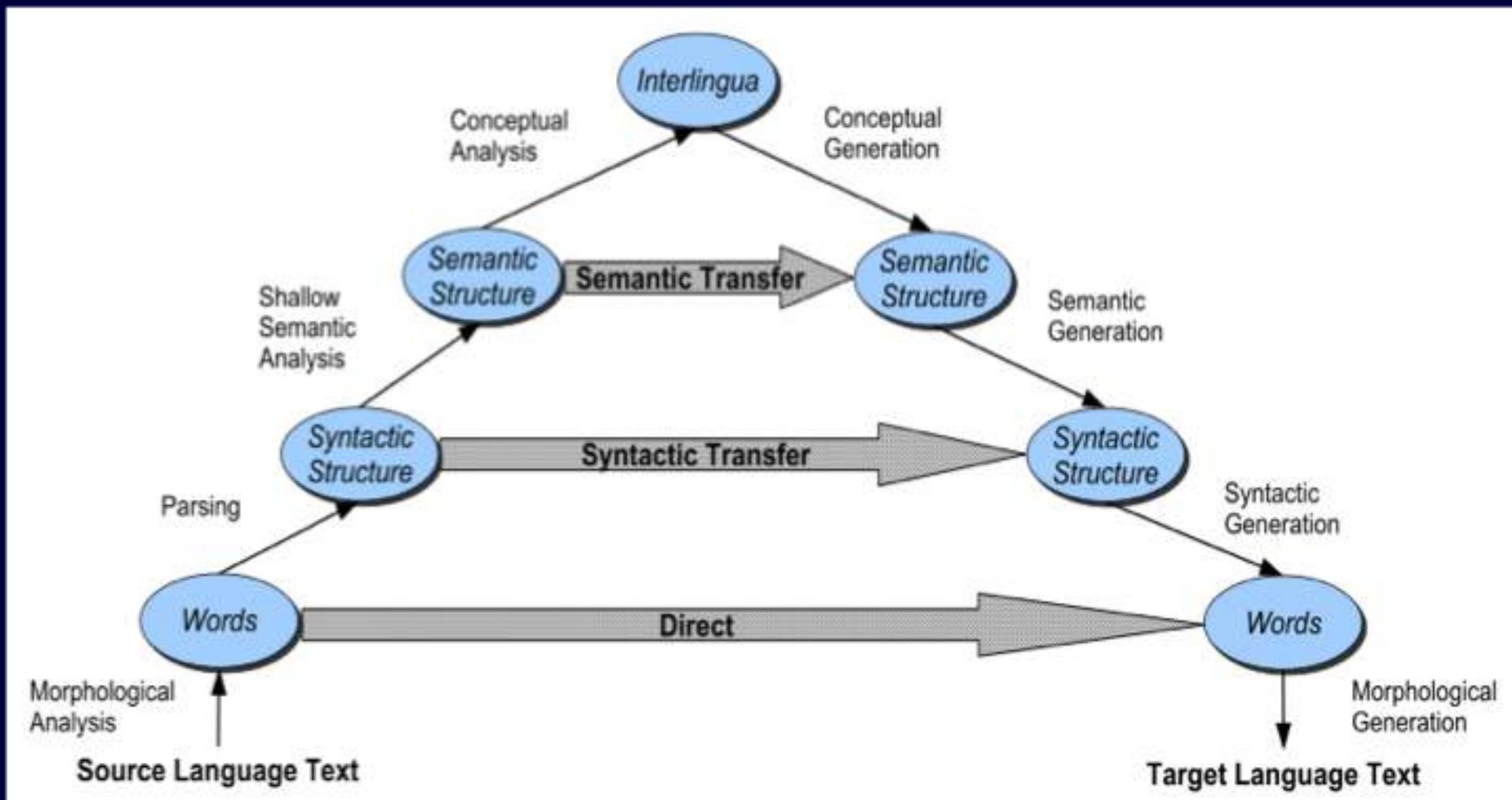
Brief History

- 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

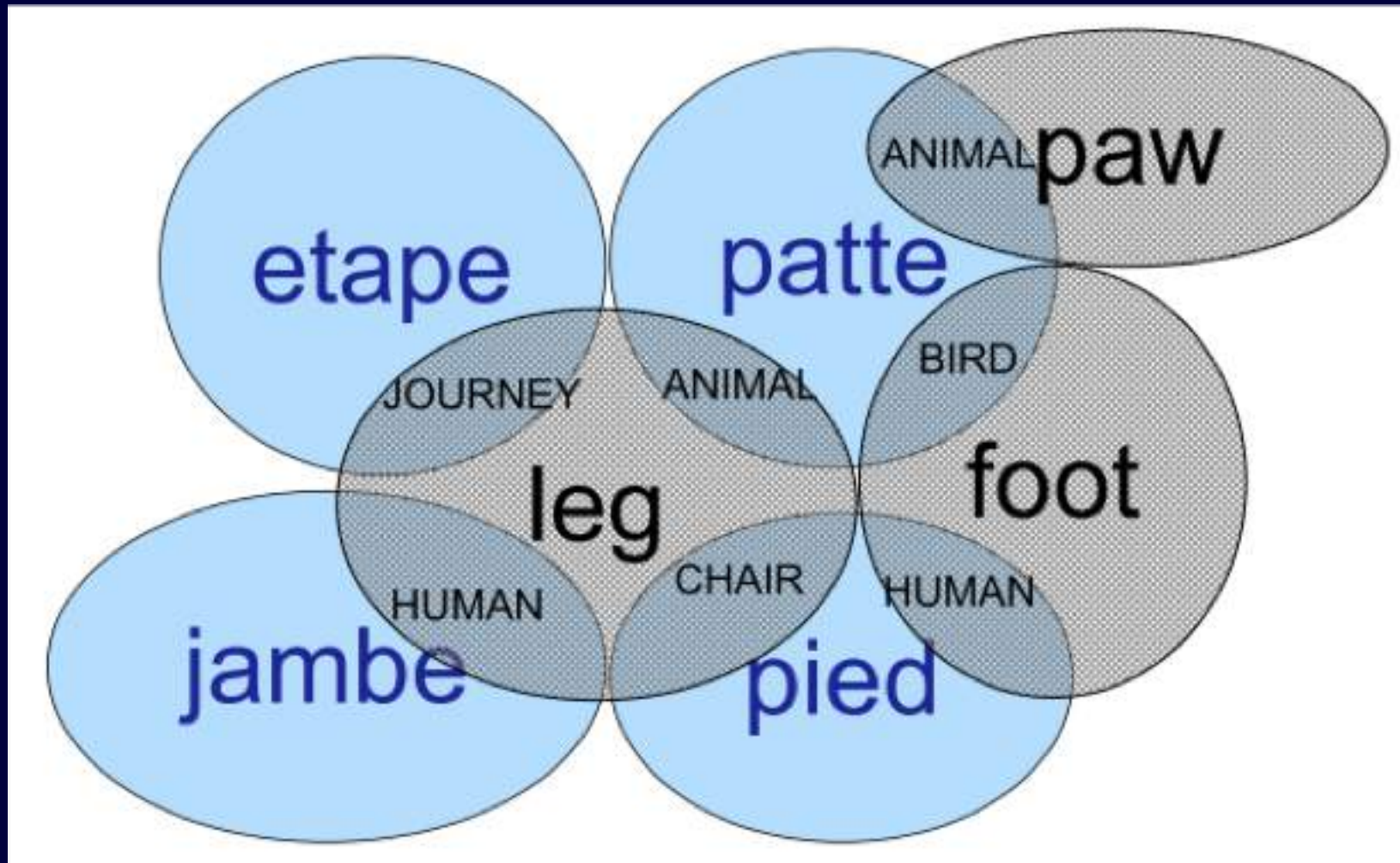
- Languages differ along many dimensions
 - Concept – Lexicon alignment – Lexical Divergence
 - Syntax – Structure Divergence
 - Word-order differences
 - English is **Subject-Verb-Object**
 - Arabic is **Verb-Subject-Object**
 - Turkish is **Subject-Object-Verb**
 - Phrase order differences
 - Structure-Semantics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

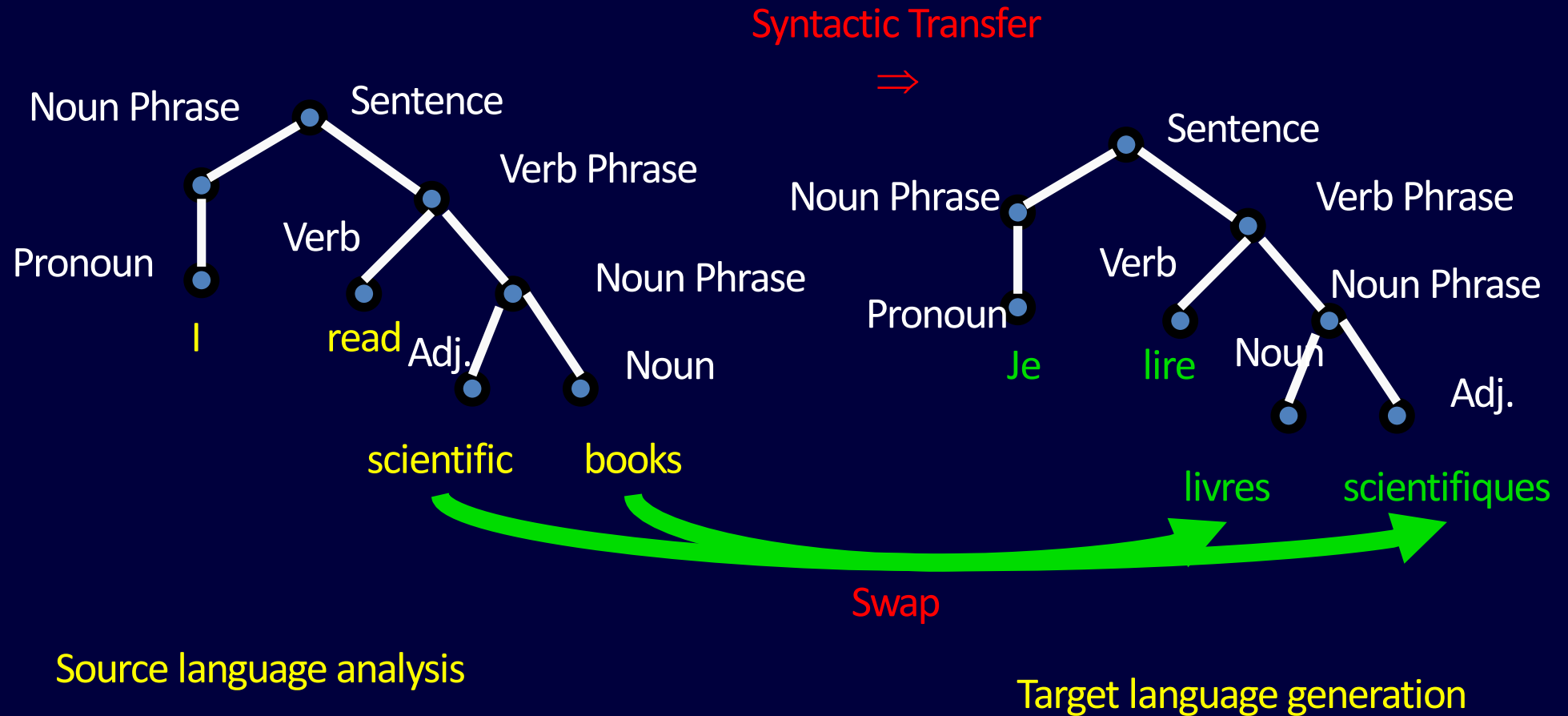
- **PANGLOSS**

- 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

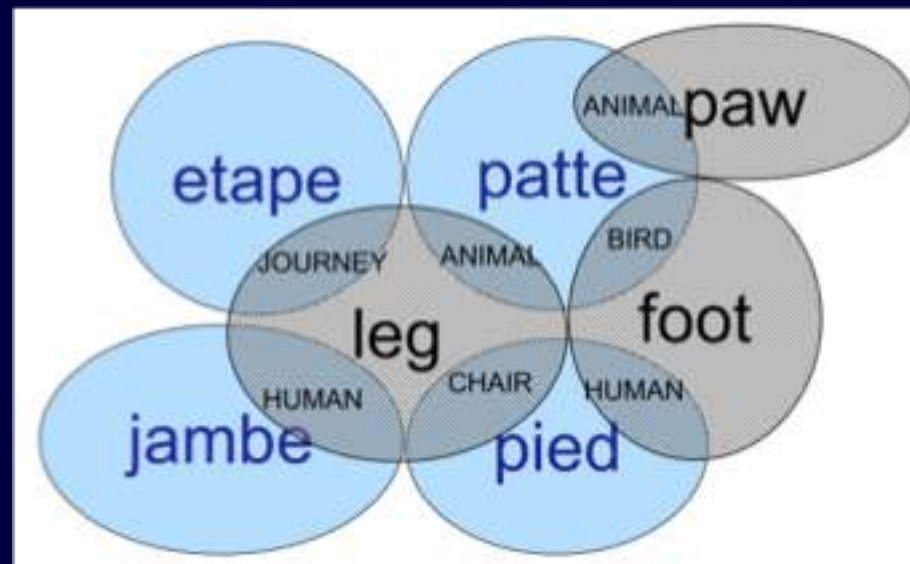


Rules

- 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: **book** (N) => French: **livre** (N, masculine)
 - English: **pound** (N, monetary sense) => French: **livre** (N, feminine)
 - English: **book** (V) => French: **réserver** (V)
- Quite tricky for



Rules

- Structure Transfer Rules

- English: $S \Rightarrow NP VP \rightarrow$

- French: $TR(S) \Rightarrow TR(NP) TR(VP)$

- English: $NP \Rightarrow Adj Noun \rightarrow$

- French: $TR(NP) \Rightarrow Tr(Noun) Tr(Adj)$

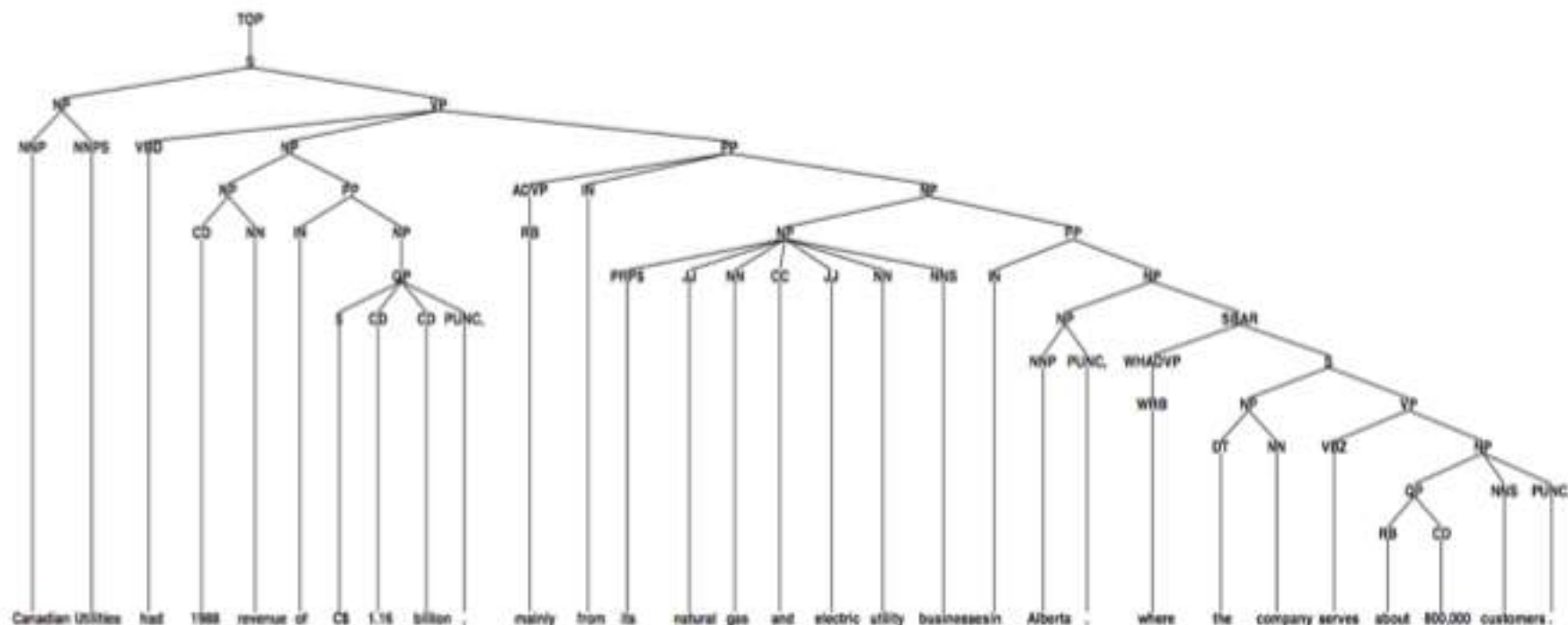
- but there are **exceptions** for

- 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)

English

- How much is that red umbrella?
- How much is that small camera?
- How much is that X?

Japanese

- Ano akai kasa wa ikura desu ka?
- Ano chiisai kamera wa ikura desu ka?
- Ano X wa ikura desu ka?

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?

Je suis fatigué.

Tired is I.

Cookies taste good!

I am tired.

Adequacy	Fluency
5	2
1	5
5	5

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)

Shallow/ Simple

Word-based
only

Phrase tables

**Example-
based MT**

Statistical MT

Knowledge
Acquisition
Strategy

Hand-built by
experts

Hand-built by
non-experts

Learn from
annotated data

Learn from un-
annotated data

All manual

Fully automated

Original **direct**
approach

Syntactic
Constituent
Structure

Semantic
analysis

Interlingua

New **Research**
Goes Here!

Typical **transfer**
system

Classic
interlingual
system

Deep/ Complex

**Knowledge
Representation
Strategy**

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 !!!

	No Probabilities	Probabilities
Flat Structure	EBMT	SMT
Deep Structure	Transfer Interlingua	Holy Grail

- Goal: structurally rich probabilistic models

Rule-based MT vs SMT

Expert System

Experts



+



Manually coded rules

*If « ... » then ...
If « ... » then ...
.....
.....
Else*

Expert system output

*T: But where are the snows
of ~~yesteryear~~?*

S: Mais où sont les neiges d'antan?

Statistical system output

*T1 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
....*

Statistical System

Bilingual parallel corpus



+



Machine
Learning

Statistical rules

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

Data-Driven Machine Translation

Man, this is so boring.

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



Translated documents



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**

Parallel Texts

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 clients 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



Cientes no venden medicinas en Europa

1a. Garcia and associates .

1b. Garcia y asociados .

7a. the clients and the associates are enemies .

7b. los clients y los asociados son enemigos .

2a. Carlos Garcia has three associates .

2b. Carlos Garcia tiene tres asociados .

8a. the company has three groups .

8b. la empresa tiene tres grupos .

3a. his associates are not strong .

3b. sus asociados no son fuertes .

9a. its groups are in Europe .

9b. sus grupos estan en Europa .

4a. Garcia has a company also .

4b. Garcia tambien tiene una empresa .

10a. the modern groups sell strong pharmaceuticals .

10b. los grupos modernos venden medicinas fuertes .

5a. its clients are angry .

5b. sus clientes estan enfadados .

11a. the groups do not sell zenzanine .

11b. los grupos no venden zanzanina .

6a. the associates are also angry .

6b. los asociados tambien estan enfadados .

12a. the small groups are not modern .

12b. los grupos pequenos no son modernos .

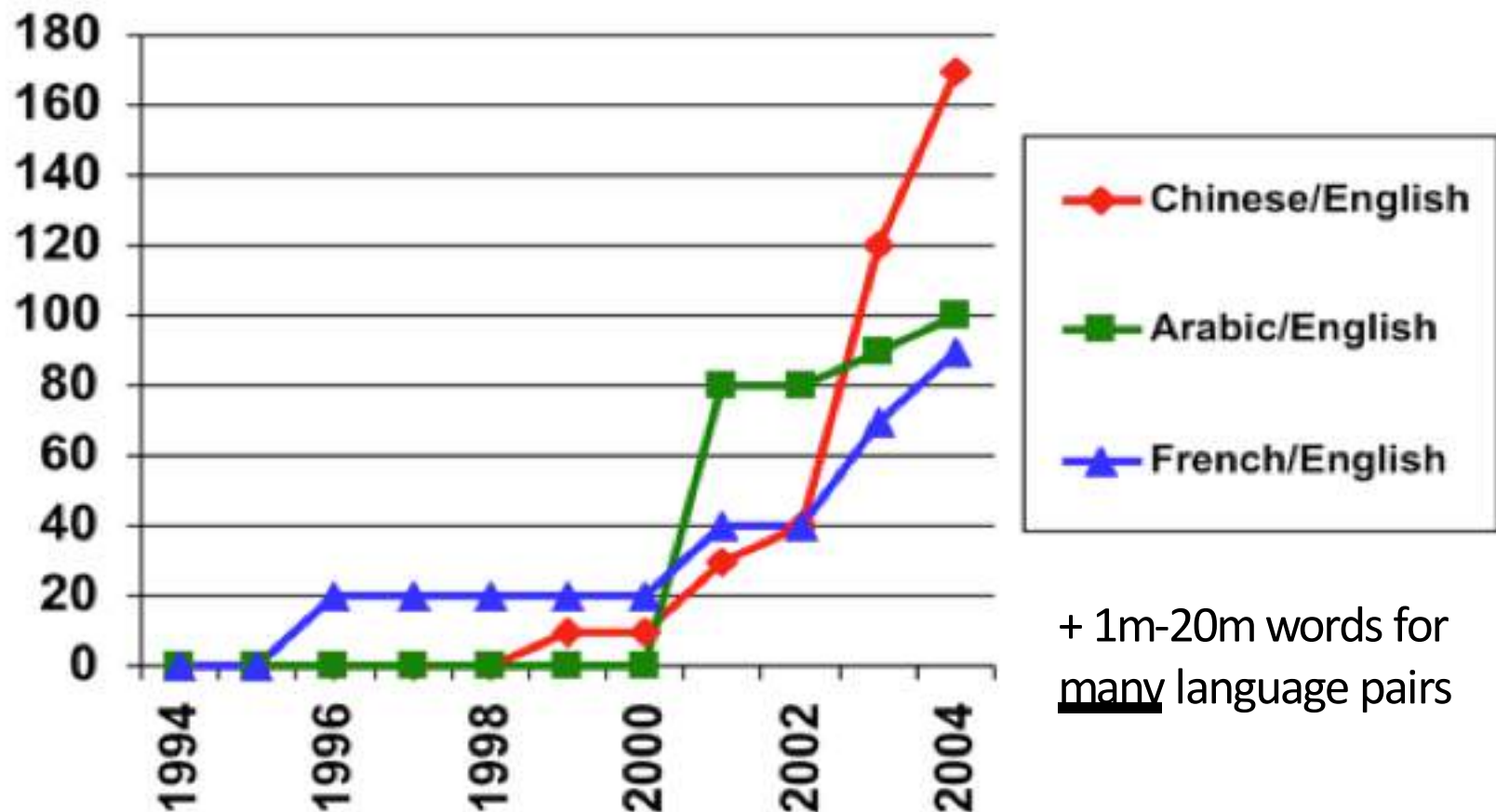
Parallel Texts

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 çoğunda 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)

Millions of
words
(English side)



+ 1m-20m words for
many language pairs

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

Available Parallel Data (2008)

- **Europarl**: 30 million words in 11 languages
- **Acquis Communautaire**: 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 open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

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.au.se>

Search & download resources:

Latest News

- 2016-01-08: New version: [OpenSubtitles2016](#)
- 2015-10-15: New versions of TED2013, NCv9
- 2014-10-24: New: JRC-Acquis
- 2014-10-20: NCv9, TED sika, DOT, WMT
- 2014-08-21: New: Ubuntu, GNOME
- 2014-07-30: New: Translated Books
- 2014-07-27: New: DOOC, Tanzil
- 2014-05-07: Parallel corpus ParCor

Search & Browse

- OPUS multilingual search interface
- Europarl v7 search interface
- Europarl v3 search interface
- OpenSubtitles search interface
- EUconst search interface
- Word Alignment Database

Sub-corpora (downloads & info):

- Books - A collection of translated literature ([DOOC2014-07-17.tar.gz](#) - 236 MB)
- DGT - A collection of EU Translation Memories provided by the JRC
- DOOC - Documents from the Catalan Government ([DOOC2014-07-17.tar.gz](#) - 702 MB)
- ECB - European Central Bank corpus
- EMA - European Medicines Agency documents ([EMEA0.3.tar.gz](#) - 5.0 GB)
- The EU bookshop corpus ([EUbookshop/ELbookshop0.2.tar.gz](#) - 33 GB)
- EUconst - The European constitution ([EUconst0.1.tar.gz](#) - 67 MB)
- EUROPARL v7 - European Parliament Proceedings ([Europarl7.tar.gz](#) - 8.4 GB)
- EUROPARL - European Parliament Proceedings ([Europarl3.tar.gz](#) - 3.6 GB)
- GNOME - GNOME localization files ([GNOME2014-06-20.tar.gz](#) - 9 GB)
- Global Voices - News stories in various languages ([GlobalVoices2013.tar.gz](#) - 1.1 GB)
- The Croatian - English WaC corpus ([HrvWaC1.tar.gz](#) - 48 MB)
- JRC-Acquis - legislative EU texts
- KDE4 - KDE4 localization files (v.2) ([KDE4.tar.gz](#) - 1.4 GB)
- KDEdoc - the KDE manual corpus ([KDEdoc.tar.gz](#) - 35 MB)
- MBS - Belgisch Staatsblad corpus
- MultiUN - Translated UN documents
- News Commentary (News-Commentary9.tar.gz - 2.2 GB)
- News Commentary (News-Commentary11.tar.gz - 741 MB)
- OO - the OpenOffice.org corpus ([OpenOffice.tar.gz](#) - 34 MB)
- OtsaPublik - Breton - French parallel texts ([OtsaPublik0.1.tar.gz](#) - 19MB)
- OpenOffice.org 3 corpus
- OpenSubtitles - the opensubtitles.org corpus
- OpenSubtitles2011 - opensubtitles.org 2011
- OpenSubtitles2012 - opensubtitles.org 2012
- OpenSubtitles2013 - opensubtitles.org 2013 (extending OpenSubtitles2012)
- OpenSubtitles2016 - opensubtitles.org 2016 (including all previous data files)
- PHP - the PHP manual corpus ([PHP.tar.gz](#) - 172 MB)
- ParCor - A Parallel Primoun-Coeference Corpus
- Regeringsförklaringen - a tiny example corpus
- SETIMES - A parallel corpus of the Balkan languages ([SETIMES1.tar.gz](#) - 2.3 GB)
- SETIMES2 - A new version of SETIMES ([SETIMES2.tar.gz](#) - 2.9 GB)
- SPC - Stockholm Parallel Corpora ([SPCv1.tar.gz](#) - 3.5 MB)
- Tatoeba - A DB of translated sentences ([Tatoeba2014-07-28.tar.gz](#) - 262MB MB)
- TedTalks fr-en ([TedTalks0.1.tar.gz](#) - 26 MB)
- TED Talks (TED2013v1) [tar.gz](#) - 781 MB)
- Tanzil - A collection of Quran translations
- TEP - The Tehran English-Persian subtitle corpus ([TEP0.1.tar.gz](#) - 49 MB)
- Ubuntu - Ubuntu localization files ([Ubuntu14.10.tar.gz](#) - 1.3 GB)
- UN - Translated UN documents ([UN20090831.tar.gz](#) - 208MB)
- WikiSource (small en-sv sample only)
- Wikipedia - translated sentences from Wikipedia ([Wikipedia1.0.tar.gz](#) - 7.8GB)
- WMT News Test Sets ([WMT-News1.tar.gz](#) - 34MB)

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 treebanks (demo)
- Uplink at bitbucket
- A reliable Language Identifier
- Scripts for OpenSubtitles2012/2013

Some Projects using OPUS

- Let'sMT! - On-line SMT toolkit
- CASMACAT - Computer-Aided Translation
- WMT - A conference on statistical MT
- Reverso - Translations in context
- SketchEngine - Tools for lexicographers
- sub-a-sub - Translations in colloquial language

Links to other Resources

- The EuroParl corpus and WMT data
- CoStEP - A cleaner and structured version of the Europarl corpus
- JRC-Acquis and related resources
- Parallel corpora at PELCRA (word-aligned data)
- UM - a domain specific Chinese-English parallel corpus
- Let's MT! and its Resource Repository Software
- Links to alignment and MT-related tools
- Links to other MT-related resources

en-20m words for
any 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 the 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 governorates

v.2.0 – October 2003

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.

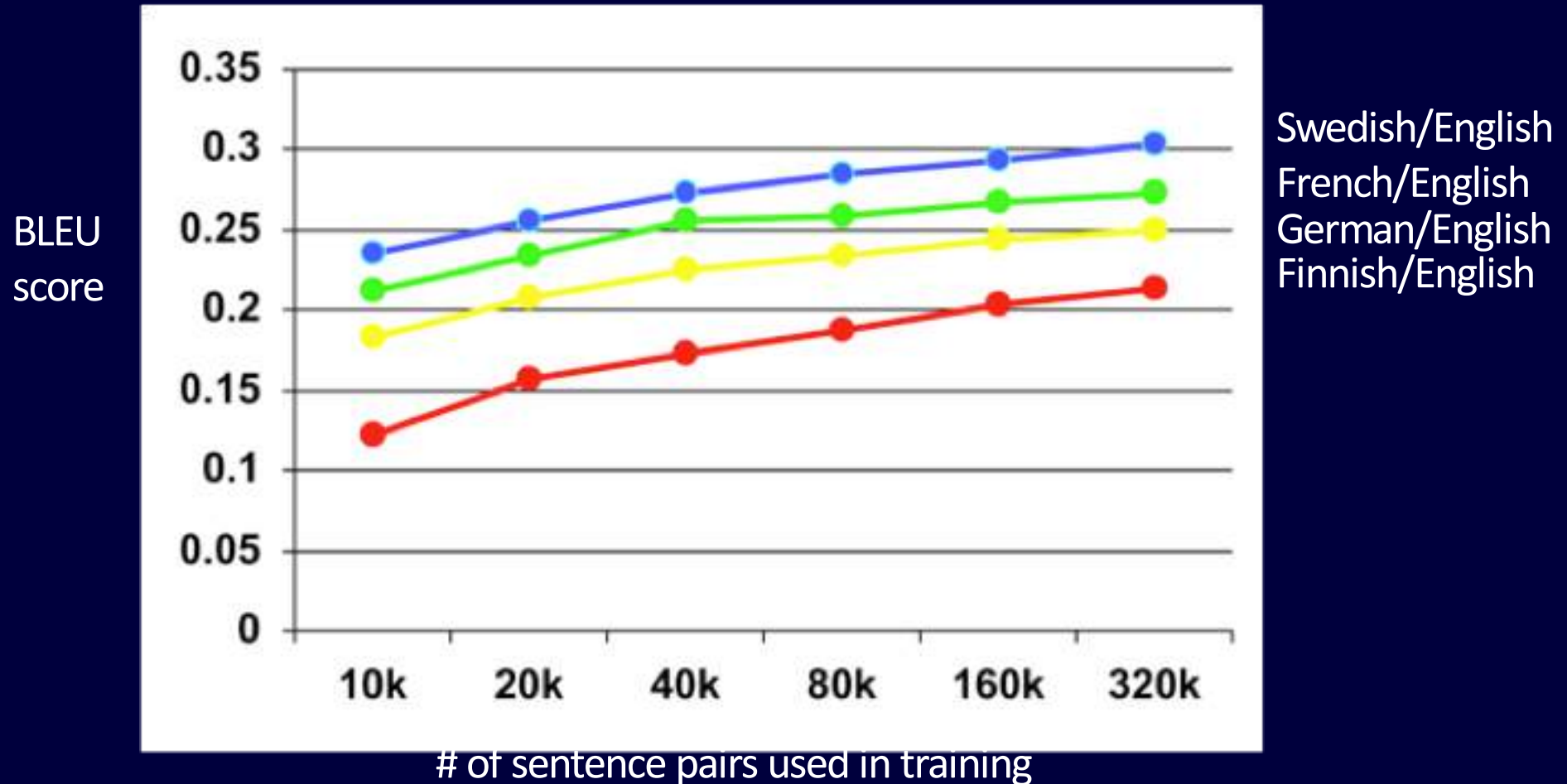
Bush said in press statements that it is possible to organize elections in most regions of Iraq to the deadline and I wish that the turnout are high. He added that "14 governorates of Iraq's 18 appeared in relative calm".

v.2.4 – October 2004

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.3.0 - February 2005

Sample Learning Curves



Experiments by
Philipp Koehn

Preparing Data

- Sentence Alignment
- Tokenization/Segmentation

Sentence Alignment

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..

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

$$p(t | s)$$

- 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

$$t^* = \operatorname{argmax}_t p(t \mid s)$$

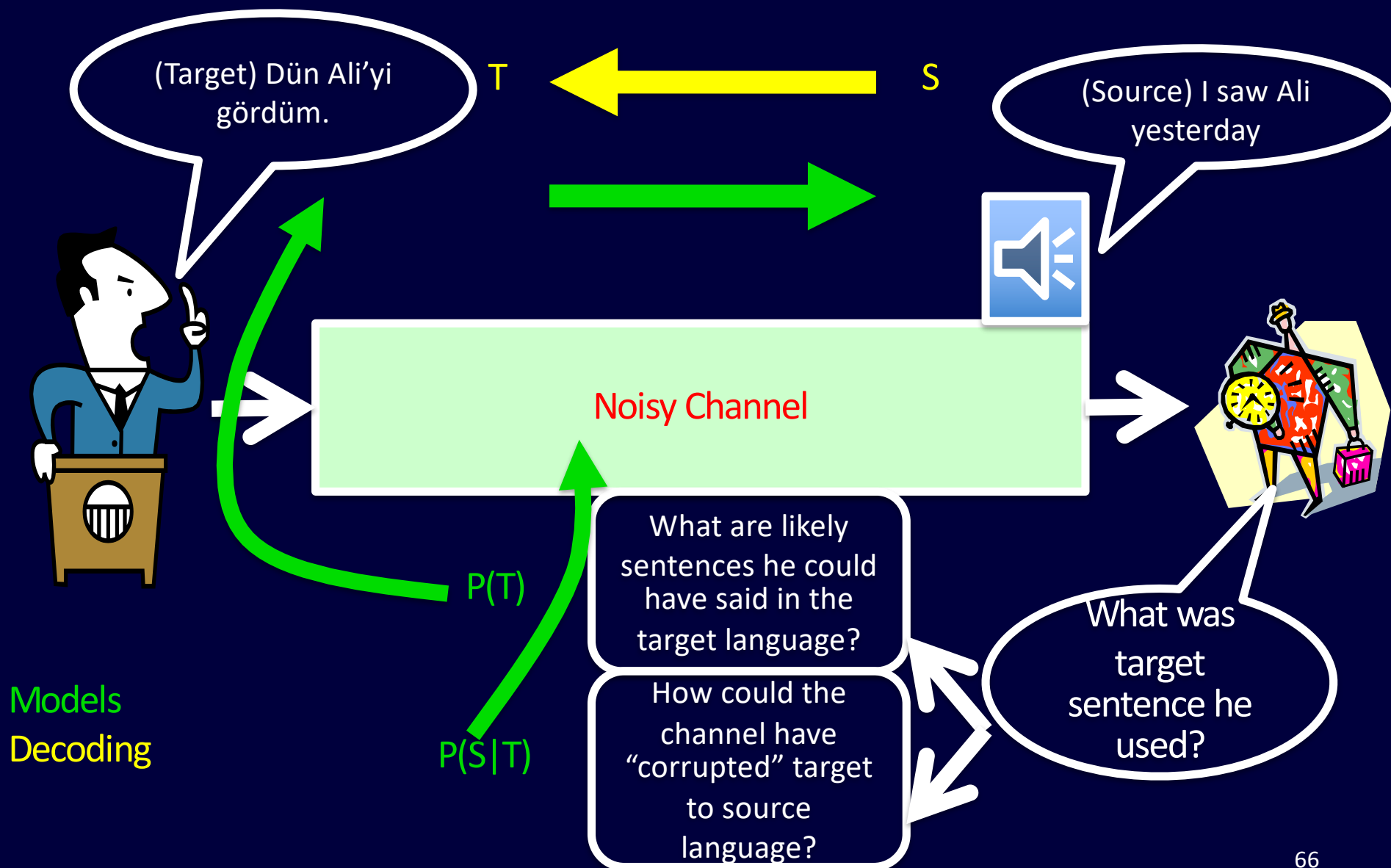
The Basic Formulation of SMT

- We need to compute $t^* = \operatorname{argmax}_t p(t | s)$
- Using Bayes' Rule we can “factorize” this into two separate problems

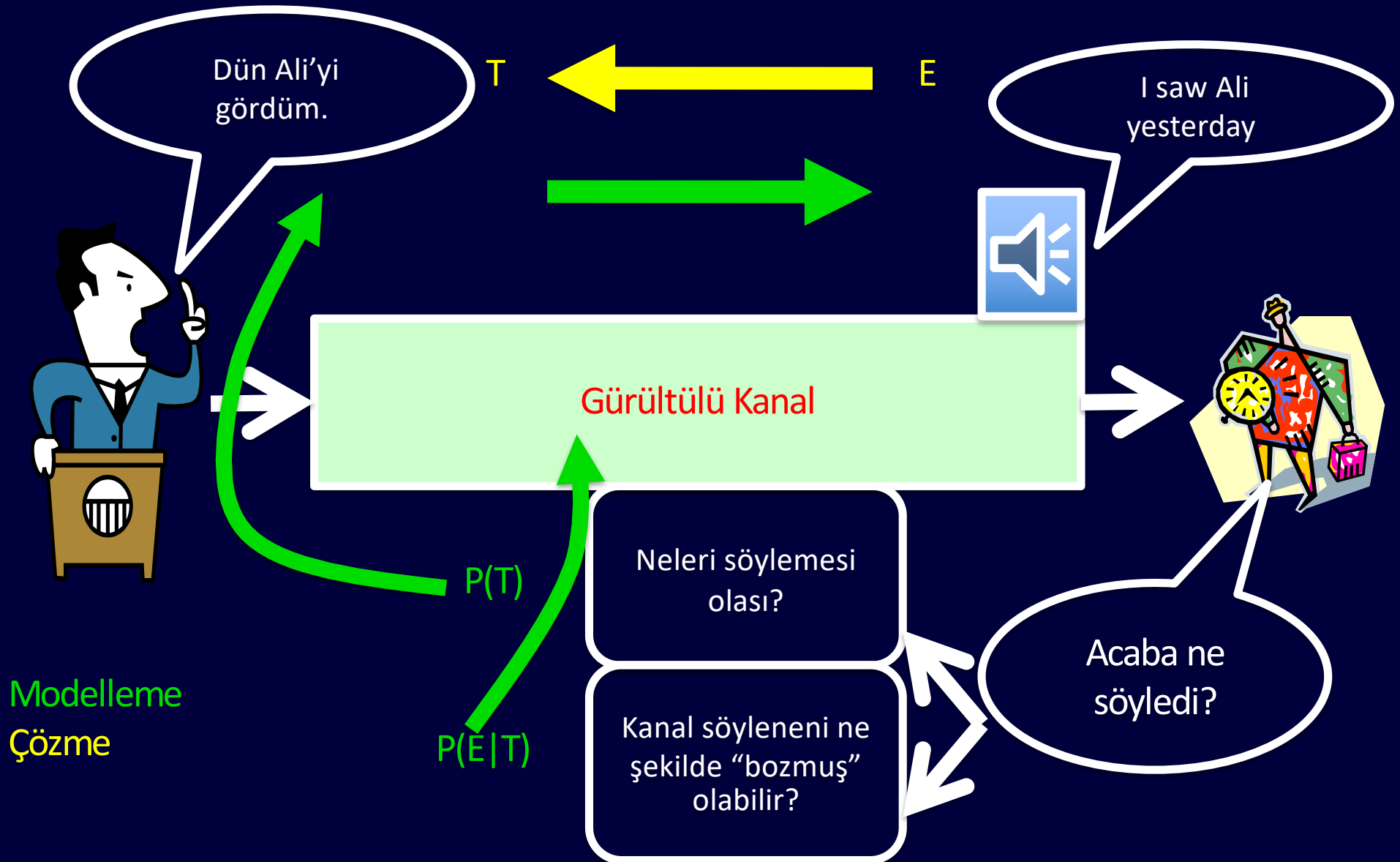
$$\begin{aligned} t^* &= \operatorname{argmax}_t \frac{p(s|t)p(t)}{p(s)} \\ &= \operatorname{argmax}_t p(s|t)p(t) \end{aligned}$$

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

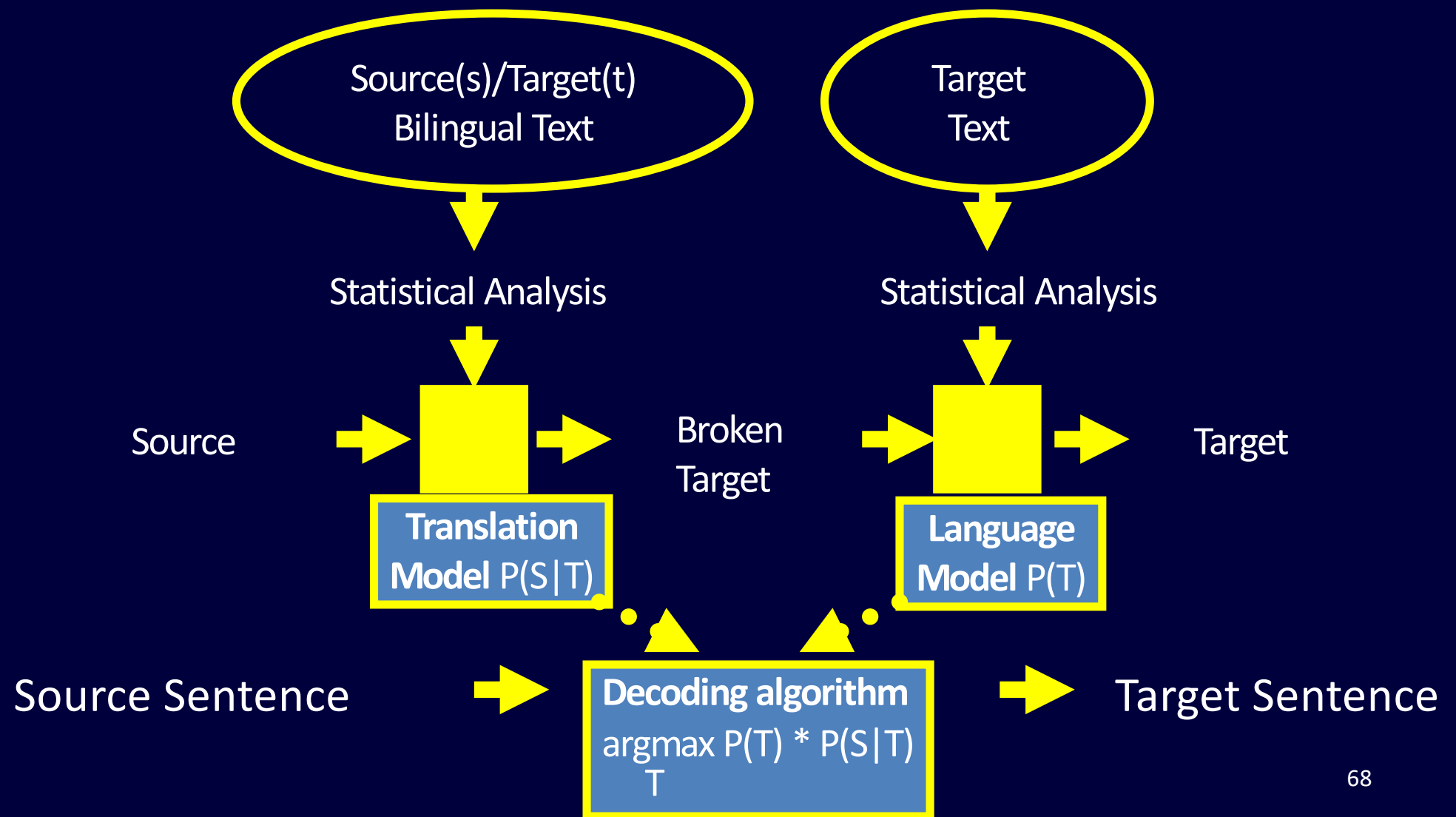
The Noisy Channel Model



The Noisy Channel Model



Where do the probabilities come from?




















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 $p(S | T) P(T)$
 - $p(T)$ models “good” target sentences (Target Language Model)
 - $p(S/T)$ models whether words in source sentence are “good” translation of words in the target sentence (Translation Model)

I saw Ali yesterday	Good Target? $P(T)$	Good match to Source ? $P(S T)$	Overall
Bugün Ali'ye gittim			
Okulda kalmışlar			
Var gelmek ben			
Dün Ali'yi gördüm			
Gördüm ben dün Ali'yi			
Dün Ali'ye gördüm			

Three Problems for Statistical MT

- **Language model**

- Given a target sentence T , assigns $p(T)$
 - good target sentence \rightarrow high $p(T)$
 - word salad \rightarrow low $p(T)$

- **Translation model**

- Given a pair of strings $\langle S, T \rangle$, assigns $p(S | T)$
 - $\langle S, T \rangle$ look like translations \rightarrow high $p(S | T)$
 - $\langle S, T \rangle$ don't look like translations \rightarrow low $p(S | T)$

- **Decoding algorithm**

- Given a language model, a translation model, and a new sentence S ... find translation T maximizing $p(T) * p(S|T)$

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)
- $T = t_1 t_2 t_3 \dots t_n$;
- We want $P(T)$ to be “high”
- A good approximation is by short n-grams
 - $P(T) \cong P(t_1 | \text{START}) \cdot P(t_2 | \text{START}, t_1) \cdot P(t_3 | t_1, t_2) \cdot \dots \cdot P(t_i | t_{i-2}, t_{i-1}) \cdot \dots \cdot P(t_n | t_{n-2}, t_{n-1})$
 - 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

$P(\text{I saw water on the table}) \cong$

$P(\text{I} \mid \text{START}) *$

$P(\text{saw} \mid \text{I}) *$

$P(\text{water} \mid \text{saw}) *$

$P(\text{on} \mid \text{water}) *$

$P(\text{the} \mid \text{on}) *$

$P(\text{table} \mid \text{the}) *$

$P(\text{END} \mid \text{table})$

The Classic Language Model

- If the target language is English, using 3-grams
 $P(\text{I saw water on the table}) \cong$

$P(\text{I} \mid \text{START, START}) *$

$P(\text{saw} \mid \text{START, I}) *$

$P(\text{water} \mid \text{I, saw}) *$

$P(\text{on} \mid \text{saw, water}) *$

$P(\text{the} \mid \text{water, on}) *$

$P(\text{table} \mid \text{on, the}) *$

$P(\text{END} \mid \text{the, table})$

Translation Model?

Generative approach:



The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

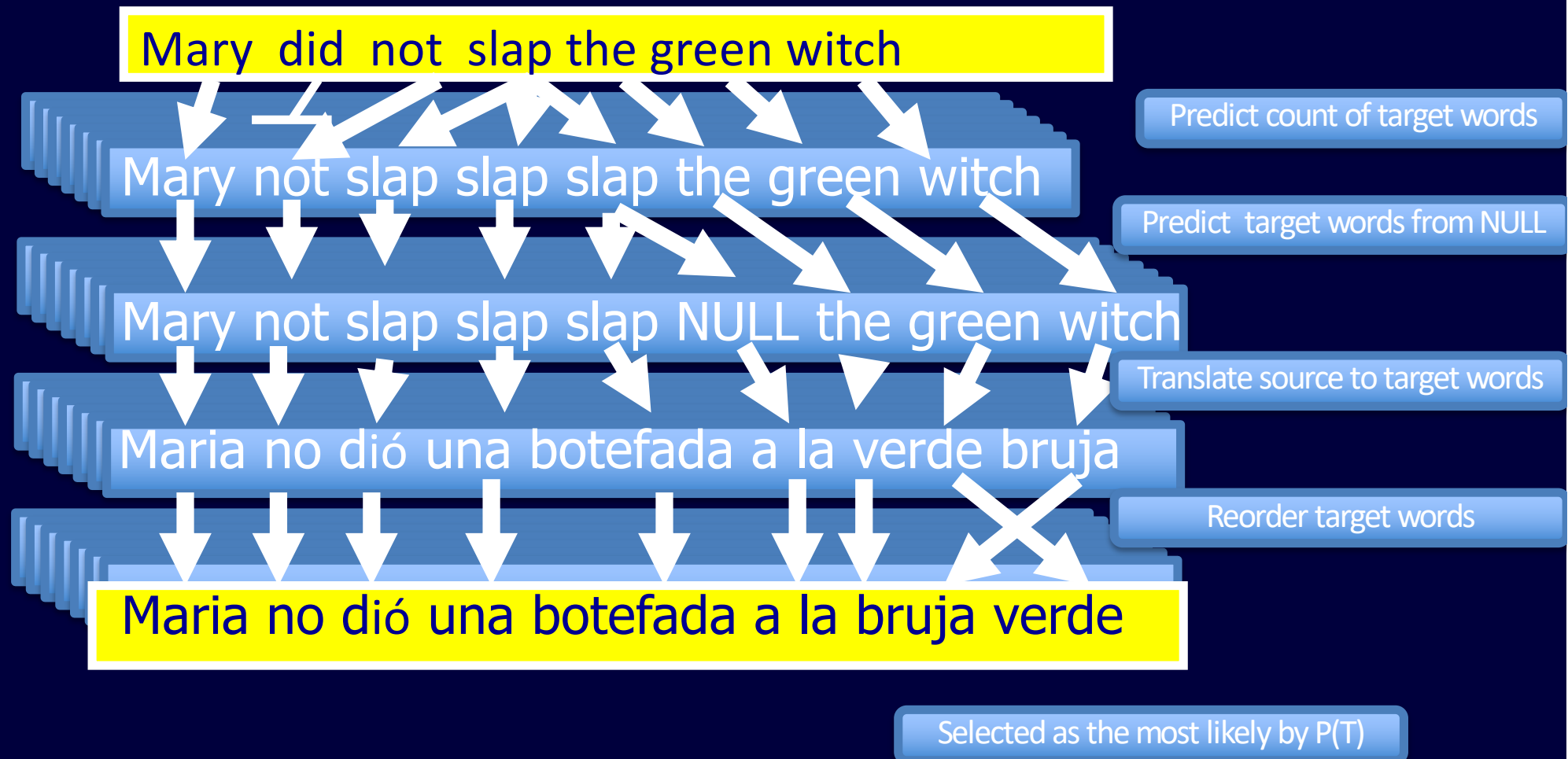
Generative approach:



The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative approach:



Basic Translation Model (IBM M-1)

- Model $p(\mathbf{t} \mid \mathbf{s}, m)$
 - $\mathbf{t} = \langle t_1, t_2, \dots, t_m \rangle, \mathbf{s} = \langle s_1, s_2, \dots, s_n \rangle$
- Lexical translation makes the following assumptions
 - Each word t_i in \mathbf{t} is generated from exactly one word in \mathbf{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 \mathbf{a} , translation decisions are conditionally independent of each other and depend only on the aligned source word t

Basic Translation Model (IBM M-1)

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



$p(\text{alignment})$



$p(\text{translation} \mid \text{alignment})$

Parameters of the IBM 3 Model

- **Fertility**: How many words does a source word get translated to?
 - $n(k | s)$: 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 $i = 1, 2, \dots, m$, choose fertility ϕ_i with probability $n(\phi_i | s_i)$.
2. Choose the number ϕ_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 $i = 0, 1, 2, \dots, m$, and each $k = 1, 2, \dots, \text{phi}_i$, choose a target word t_{ik} with probability $\text{tr}(t_{ik} \mid s_i)$.
5. For each $i = 1, 2, \dots, l$, and each $k = 1, 2, \dots, \text{phi}_i$, choose target position pi_{ik} with probability $d(pi_{ik} \mid i, l, m)$.

How IBM Model 3 works

6. For each $k = 1, 2, \dots, \phi_0$, choose a position p_{0k} from the remaining vacant positions in $1, 2, \dots, q$, for a total probability of $1/\phi_0$.
7. Output the target sentence with words t_{ik} in positions p_{ik} ($0 \leq i \leq m, 1 \leq k \leq \phi_i$).

Example

b	c	d
	+ - +	
x	y	z

b	d
x	y

- 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

b	c	d
	+ - +	
x	y	z

b	d
x	y

- t-parameters
- $t(x|b)=1.0$
- $t(y|d)=2/3$
- $t(z|d)=1/3$

Example

b	c	d
	+ - +	
x	y	z

b	d
x	y

- 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

b	c	d
	+ - +	
x	y	z

b	d
x	y

- p1
- No target words are generated by NULL so p1 = 0.0

The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative approach:



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

	Those	people	have	grown	up	,	lived	and	worked	many	years	in	a	farming	district	.
Ces	■															
gens		■														
ont			■													
grandi				■	■											
,						■										
vécu							■									
et								■								
œuvre									■							
des										■						
dizaines										■						
d'											■					
années												■				
dans													■			
le														■		
domaine															■	
agricole																■
.																■

- 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

	Those	people	have	grown	up	,	lived	and	worked	many	years	in	a	farming	district	.
Ces	■								■							
gens		■							■							
ont			■						■							
grandi				■					■							
,					■				■							
vécu						■			■							
et							■		■							
oeuvre	■	■	■	■	■	■	■	■	■	■						
des									■							
dizaines									■							
d'									■							
années									■							
dans									■							
le									■							
domaine									■							
agricole									■							
.									■							

$$tr(oeuvre | worked) = \frac{c(oeuvre | worked)}{c(worked)}$$


- Count over all aligned sentences
- worked
 - fonctionné(30), travaillé(20), marché(27), **œuvré (13)**
 - **tr(oeuvre | worked)=0.13**
- Similarly, **n(3, 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 $p(\text{french-word} \mid \text{english-word})$ 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?

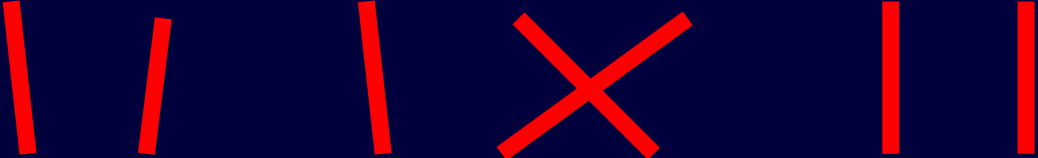
... la maison ... la maison bleue ... la fleur ...
... the house ... the blue house ... the flower ...

The diagram illustrates the process of settling down word alignments after another iteration. It shows two sentences: "... la maison ... la maison bleue ... la fleur ..." in French and "... the house ... the blue house ... the flower ..." in English. Red lines represent the current alignments: "la" to "the", "maison" to "house", "maison" to "blue", "bleue" to "house", and "fleur" to "flower". Yellow lines represent previous alignments that are being updated: "la" to "house", "maison" to "the", "maison" to "blue", "bleue" to "house", and "fleur" to "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!

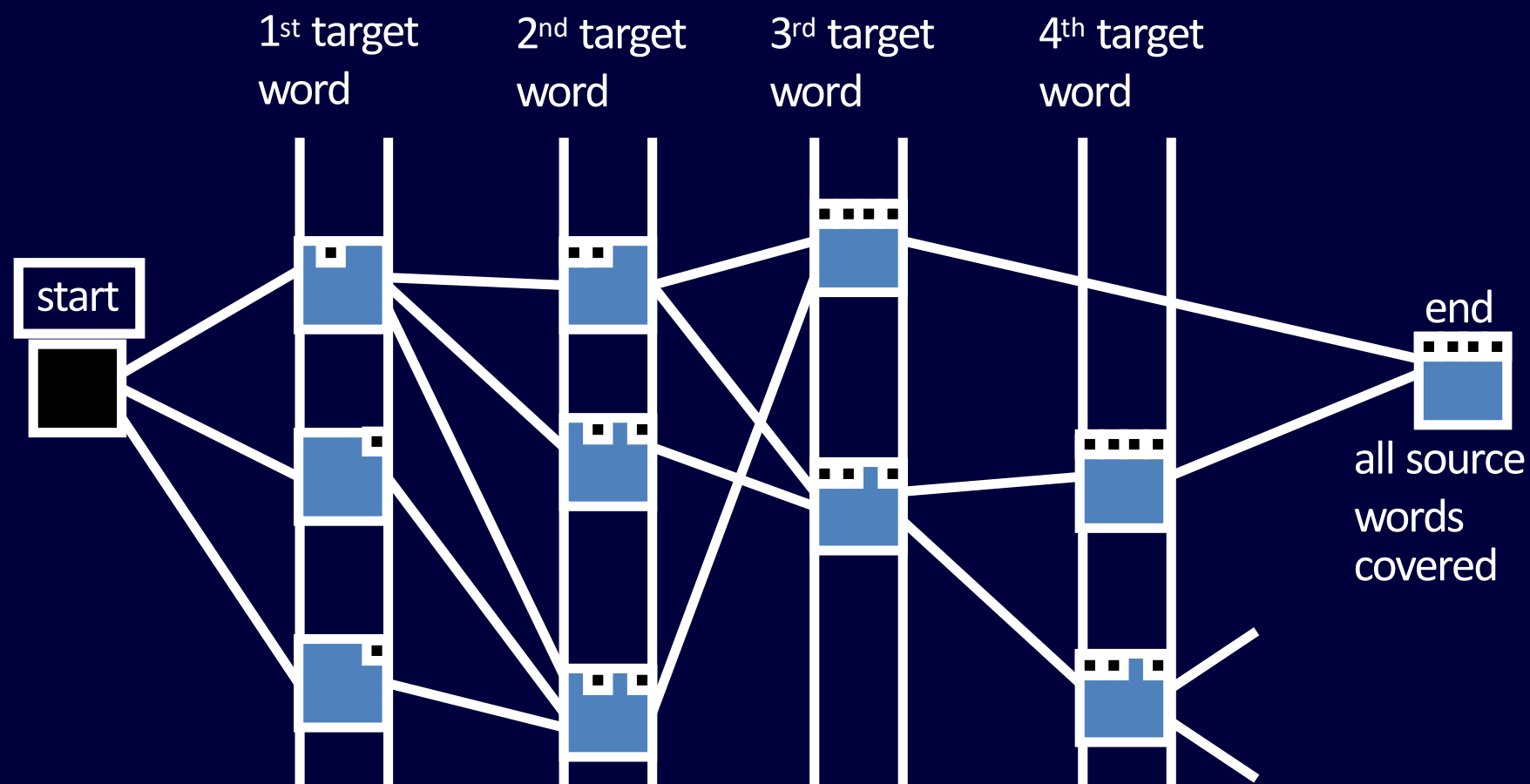
For further details, see:

- “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 $p(t) * p(t | s)$
- 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

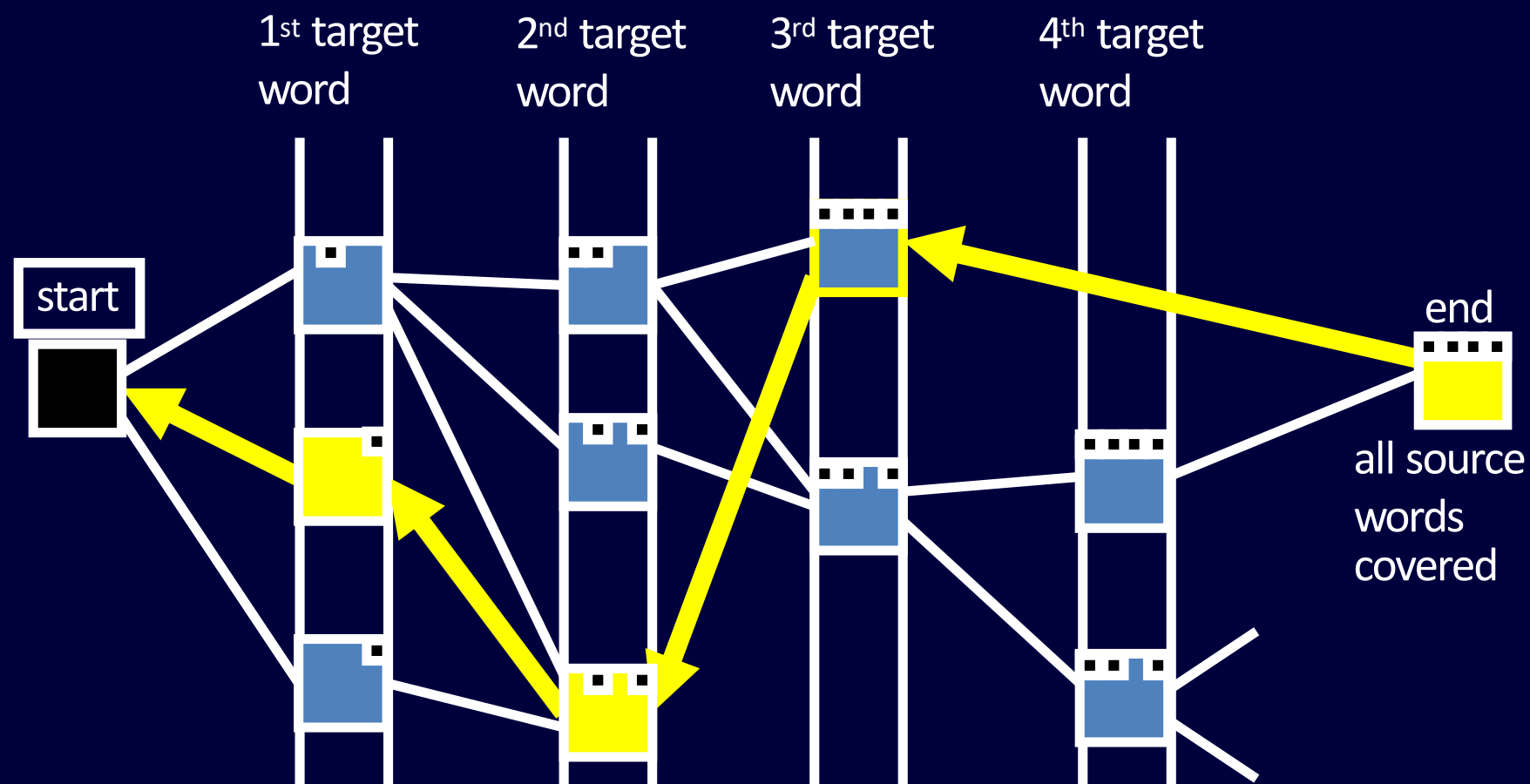


Each partial translation hypothesis contains:

- 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)

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)

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)

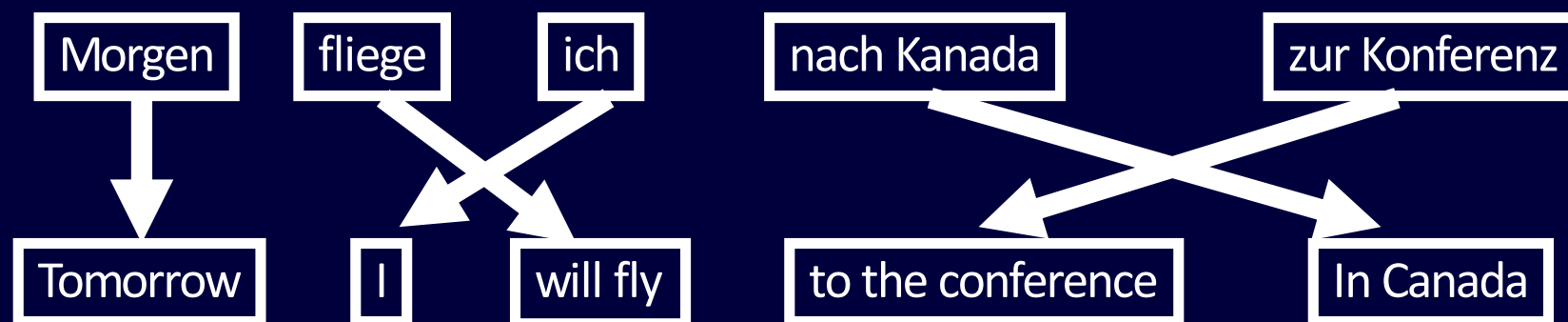
对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

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 in to 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.

	Maria	no	dió	una	bofetada	a	la	bruja	verde
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.

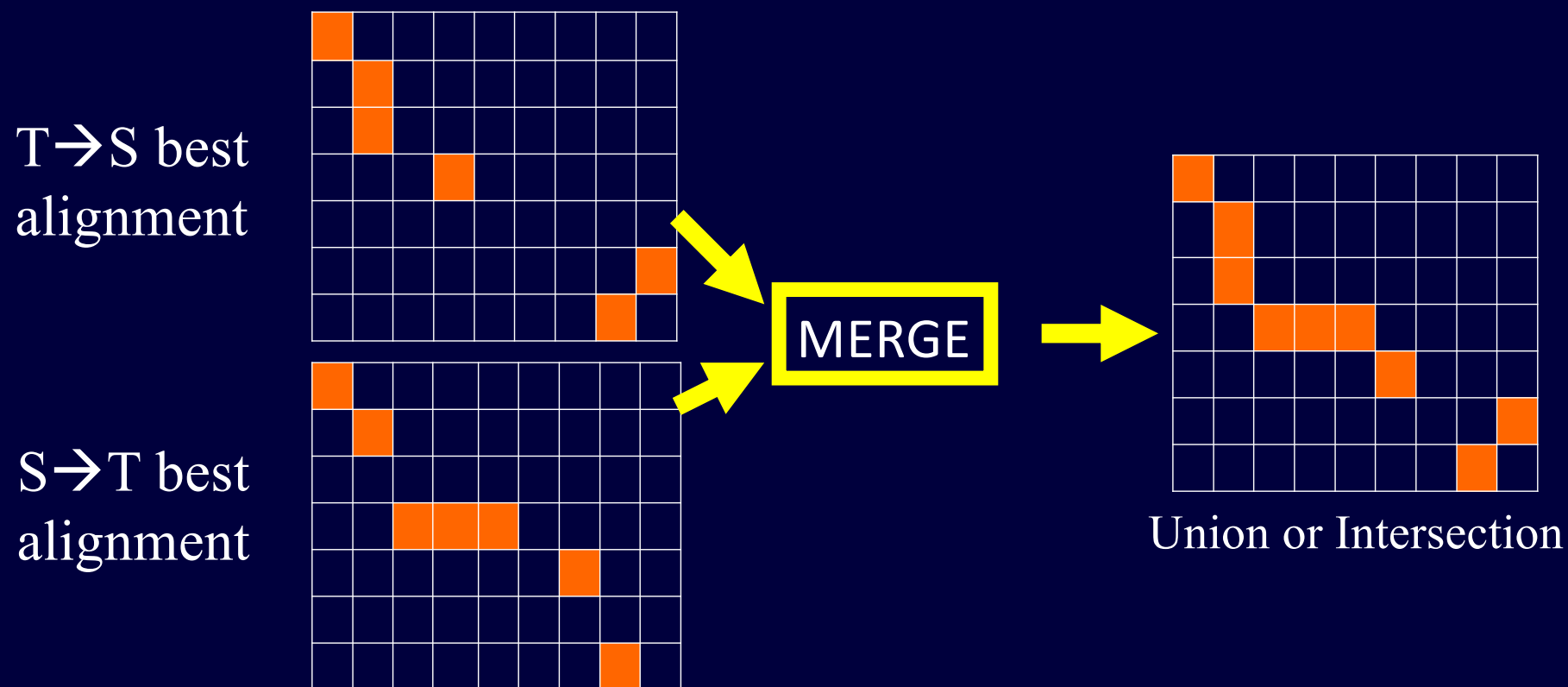
	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary									
did									
not									
slap									
the									
green									
witch									

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.

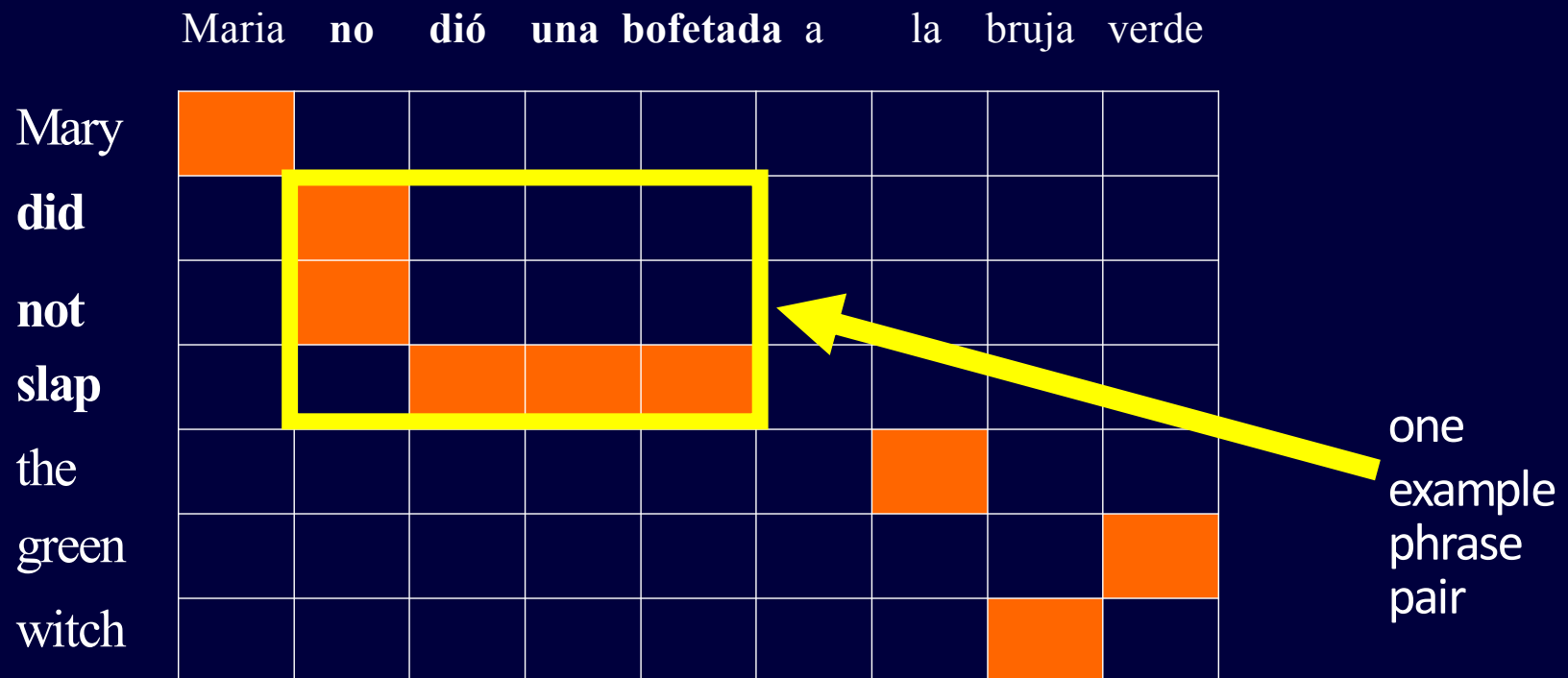
IBM Models are 1-to-Many

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



How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*



Word Alignment Consistent Phrases

	Maria	no	dió	
Mary				
did				
not				
slap				

consistent

	Maria	no	dió	
Mary				
did				
not				
slap				

inconsistent

	Maria	no	dió	
Mary				
did				
not				
slap				

inconsistent

Phrase alignment must contain all alignment points for all the words in both phrases!

Word Alignment Induced Phrases

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

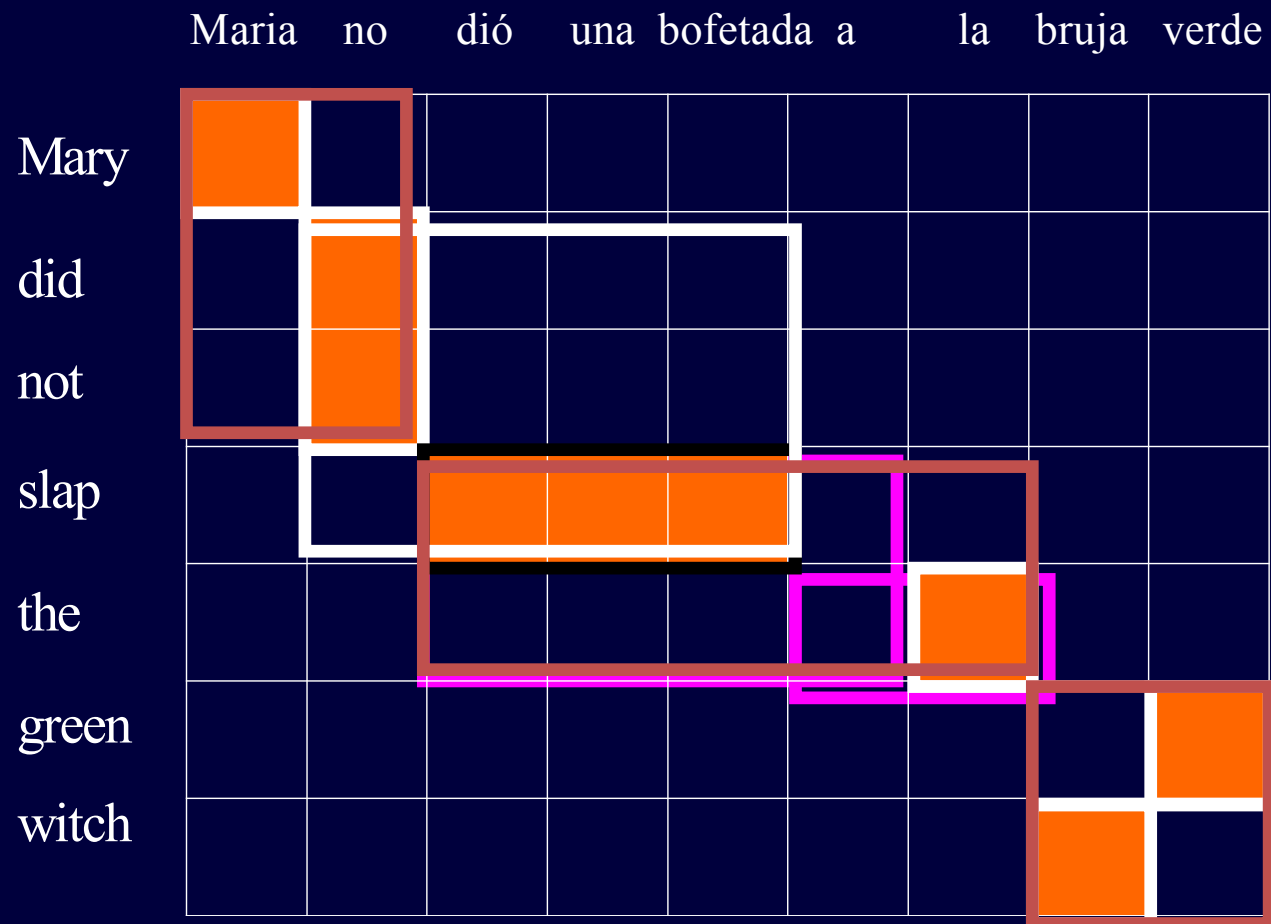
(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

Word Alignment Induced Phrases

	Maria	no	dió	una	bofetada	a	la	bruja	verde
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)

Word Alignment Induced Phrases



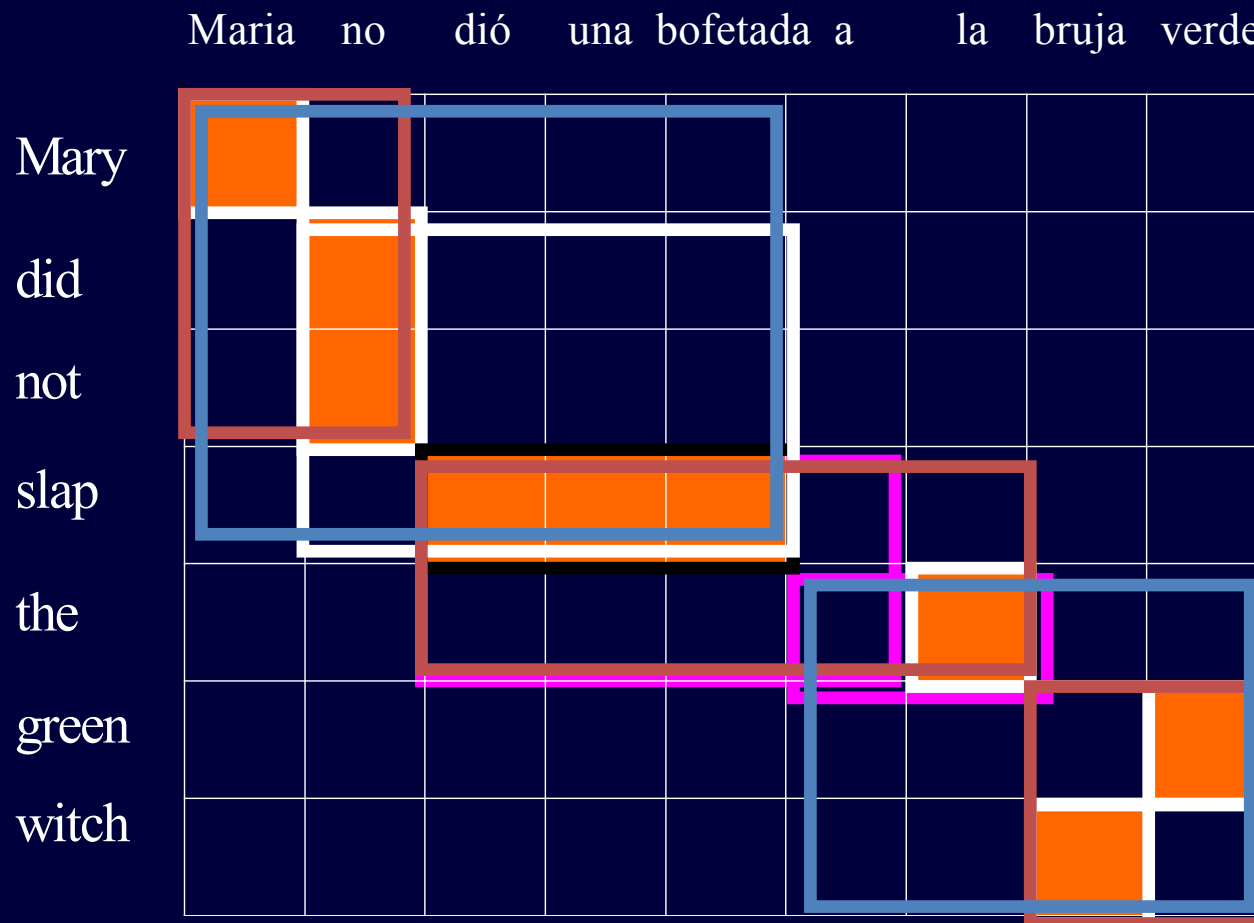
(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(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)

Word Alignment Induced Phrases



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

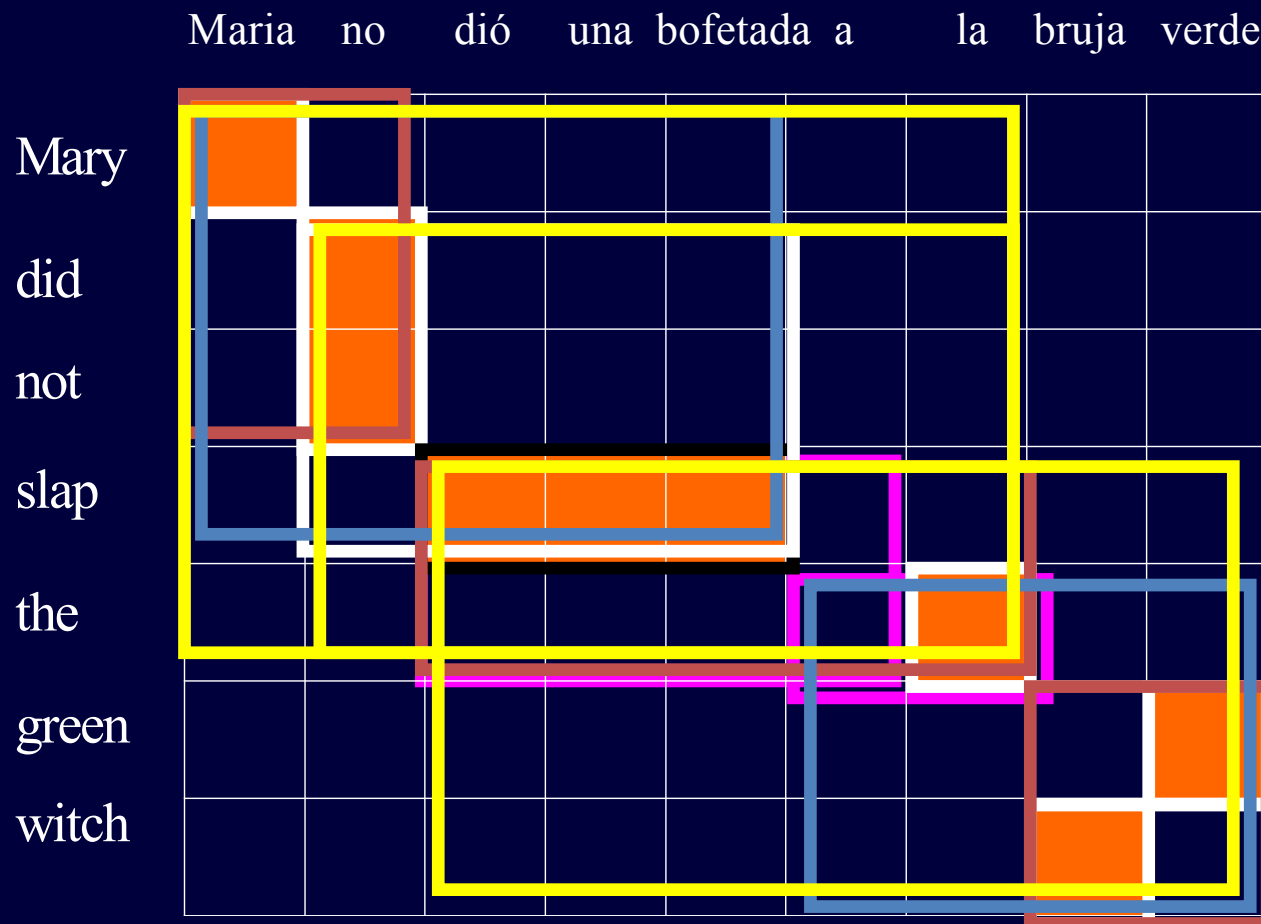
(a la, the) (dió una bofetada a, slap the)

(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)

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)

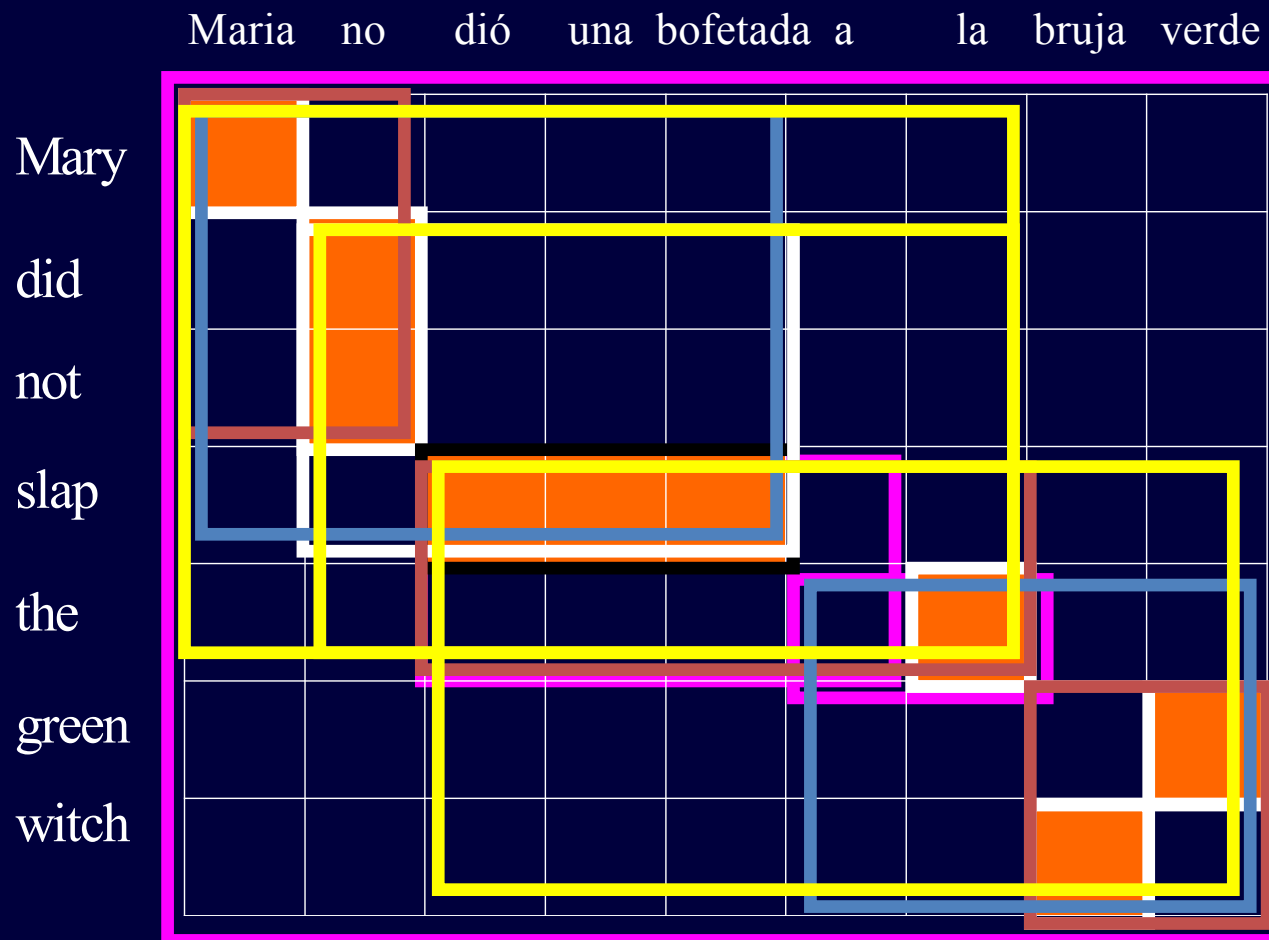
(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)

(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)

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)

(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 $p(t | s)$

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

er	geht	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home
it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	
	is		to		
	are		following		
	is after all		not after		
	does		not to		
	not				
	is not				
	are not				
	is not a				

Basic Model, Revisited

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

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

$$\operatorname{argmax}_t P(t) \times P(t \mid s)$$

Basic Model, Revisited

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

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

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

Basic Model, Revisited


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

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

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

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

Basic Model, Revisited

$$\operatorname{argmax}_e P(t)^{2.4} \times P(s \mid t) \times \text{length}(t)^{1.1} \times \text{KS}^{3.7} \dots$$


Lots of **knowledge sources** vote on any given hypothesis.

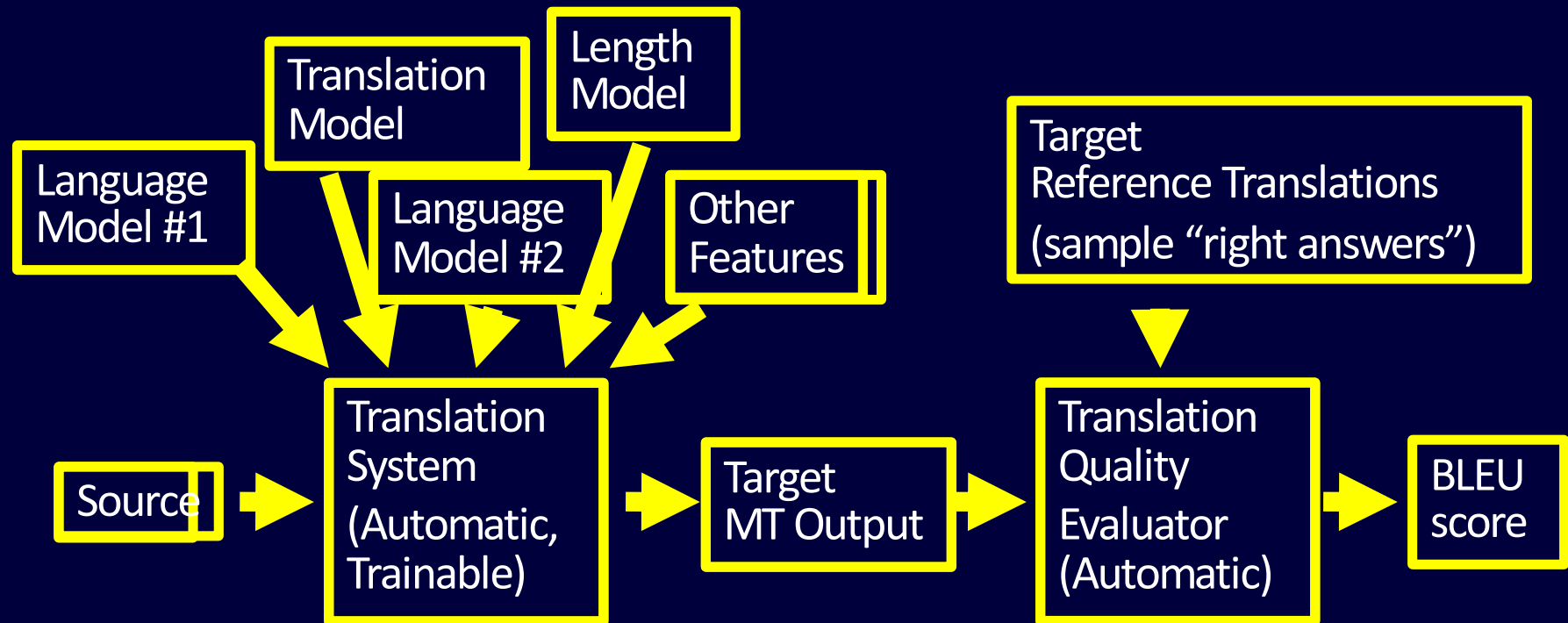
“Knowledge source” = “feature function” = “score component”.

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 a [?] 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.

More Reference Translations are Better

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 [?] receives one calls self the said 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

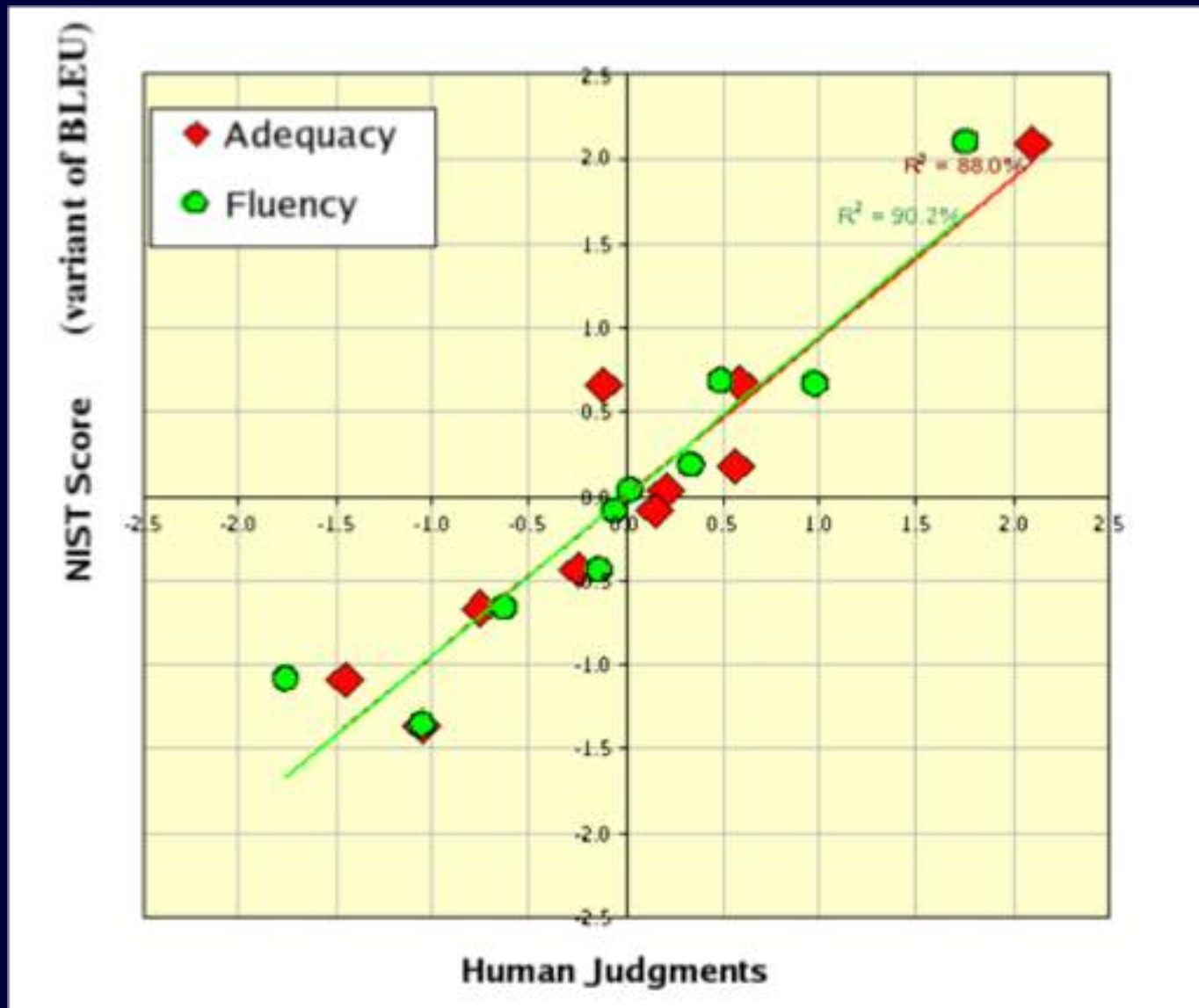
- Reference Translation: *The gunman was shot to death by the police .*
- The gunman was shot kill .
- Wounded police jaya of
- The gunman was shot dead by the police .
- The gunman arrested by police kill .
- The gunmen were killed .
- The gunman was shot to death by the police .
- The ringer is killed by the police .
- Police killed the gunman .
- Green = 4-gram match (good!) Red = unmatched word (bad!)

BLEU Formulation

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

precision_i : i-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**: wsy**ktb**unha (wa+sa+ya+**ktub**+ūn+ha “and they will **write** her”)
 - What are the alignments?
 - **Turkish**: **ge**le**bi**lecek**mi**ssin (gel+ebil+ecek+mis+sin (I heard) you would be **com**ing))
 - What are the alignments?

Morphology & SMT

- Finlandiya^lı^{la}ştı^{tır}ama^{dık}lar^ımız^{dan}mış^{sın}ız^{casına}
- Finlandiya+lı+laş+tır+ama+dık+lar+ımız+dan+mış+sını
z+casına
- (behaving) as if you have been one of those whom
we could not convert into a Finn(ish
citizen)/someone from Finland

Morphology & SMT

- yapabileceksek

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

Most of the time, the morpheme order is “reverse” of the corresponding English word order

- yaptırabildiğimizde

- yap+tır+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 clumsy+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”

Word Form	Count	Gloss
faaliyet	3	activity
faaliyete	1	to the activity
faaliyetinde	1	in its activity
faaliyetler	3	activities
faaliyetlere	6	to the activities
faaliyetleri	7	their activities
faaliyetlerin	7	of the activities
faaliyetlerinde	1	in their activities
faaliyetlerine	5	to their activities
faaliyetlerini	1	their activities (accusative)
faaliyetlerinin	2	of their activities
faaliyetleriyle	1	with their activities
faaliyette	2	in (the) activity
faaliyetteki	1	that is in activity
TOTAL	41	

An Example E – T Translation

we are going to your hotel in Taksim by taxi



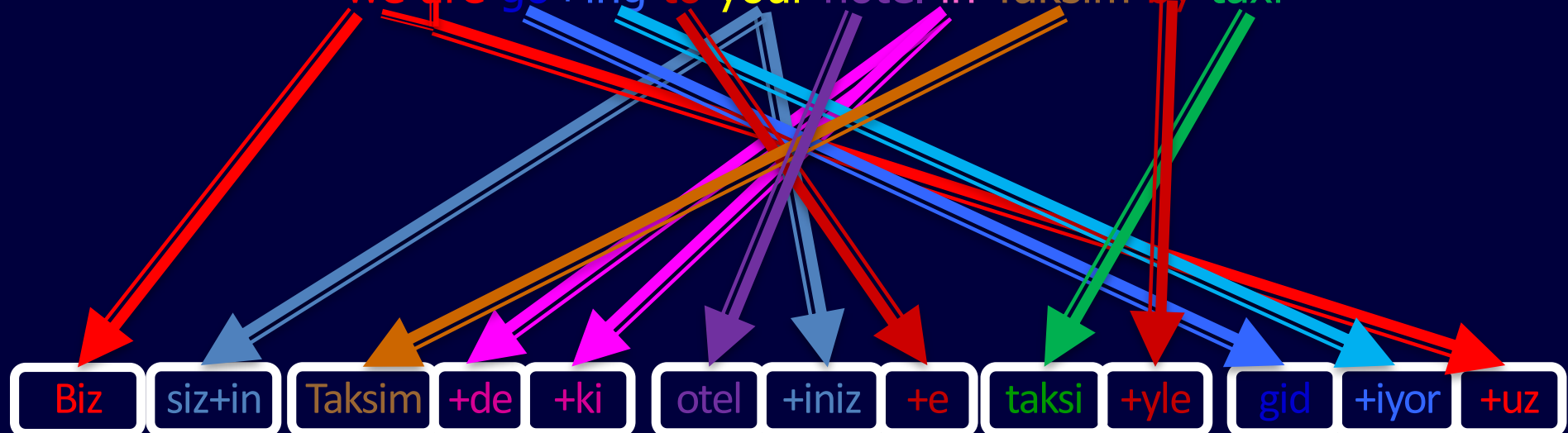
we are go+ing to your hotel in Taksim by taxi

An Example E – T Translation

we are going to your hotel in Taksim by taxi



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



An Example E – T Translation

we are going to your hotel in Taksim by taxi



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

Biz siz+in Taksim +de +ki otel +iniz +e taksi +yle gid +iyor +uz

An Example E – T Translation

we are going to your hotel in Taksim by taxi



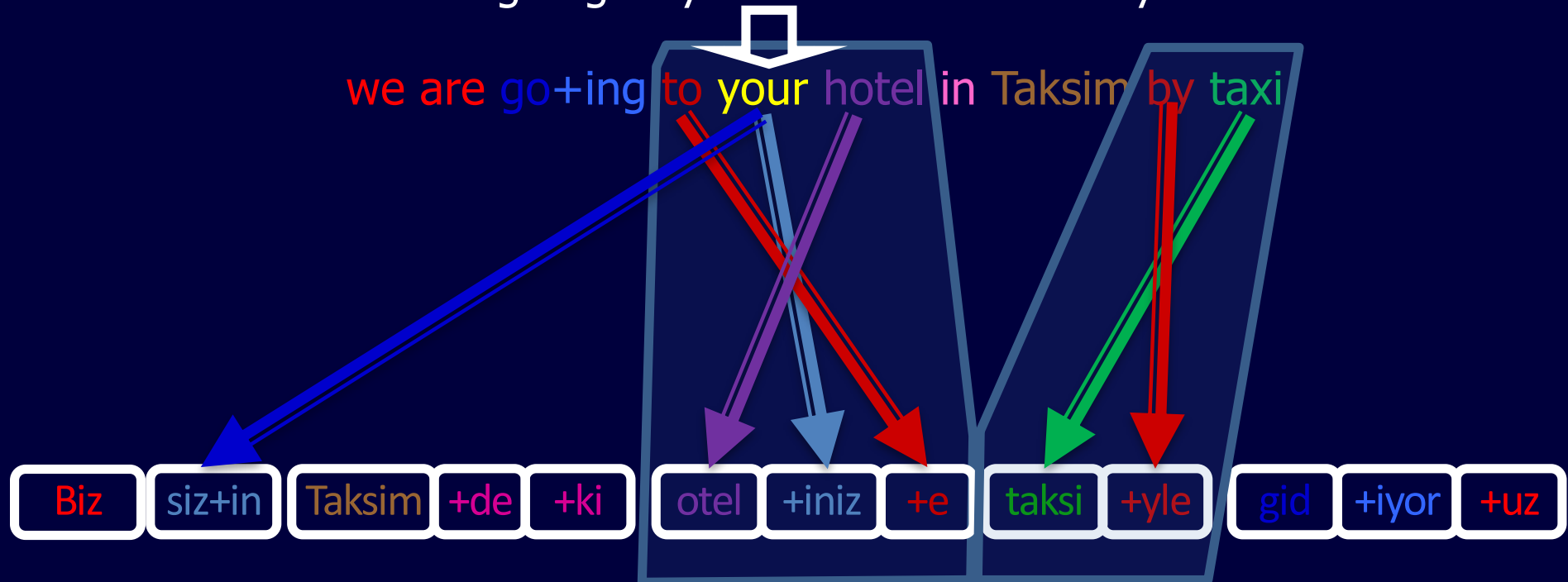
we are go+ing to your hotel in Taksim by taxi

Biz siz+in Taksim +de +ki otel +iniz +e taksi +yle gid +iyor +uz

An Example E – T Translation

we are going to your hotel in Taksim by taxi

we are go+ing 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

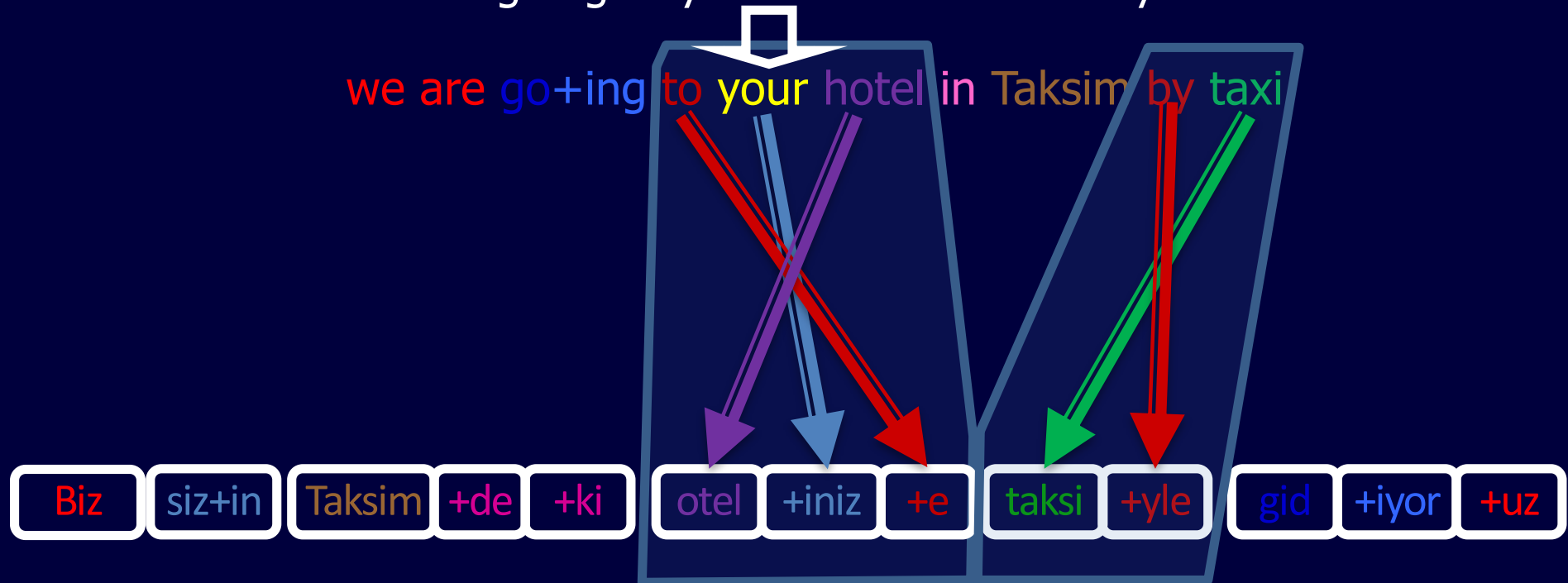
Syntax and Morphology Interaction

- 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?

Syntax and Morphology Interaction

we are going to your hotel in Taksim by taxi

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



Suppose we can do some **syntactic analysis on the English side**

Syntax and Morphology Interaction

we are go+ing 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

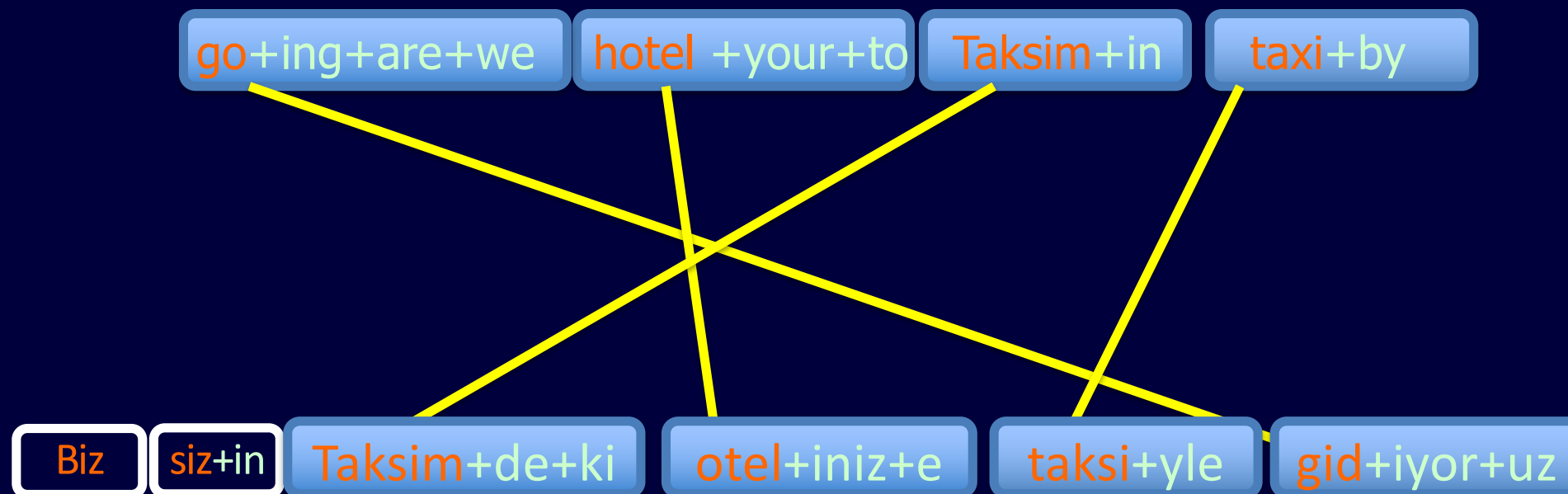
Syntax and Morphology Interaction

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

Syntax and Morphology Interaction

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

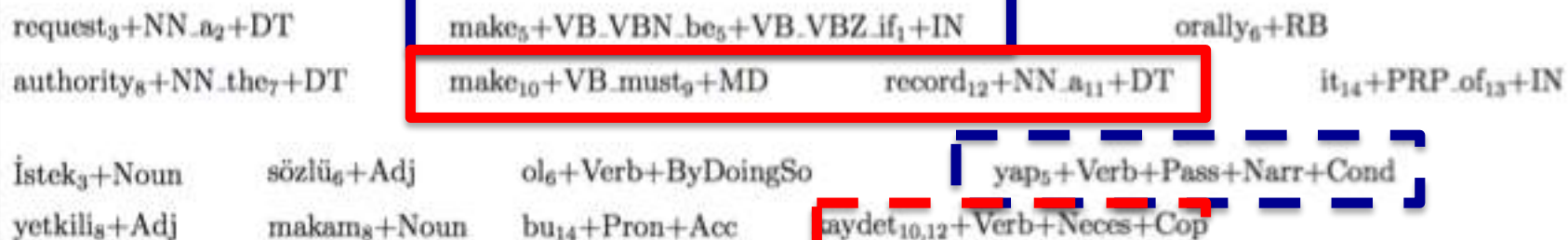
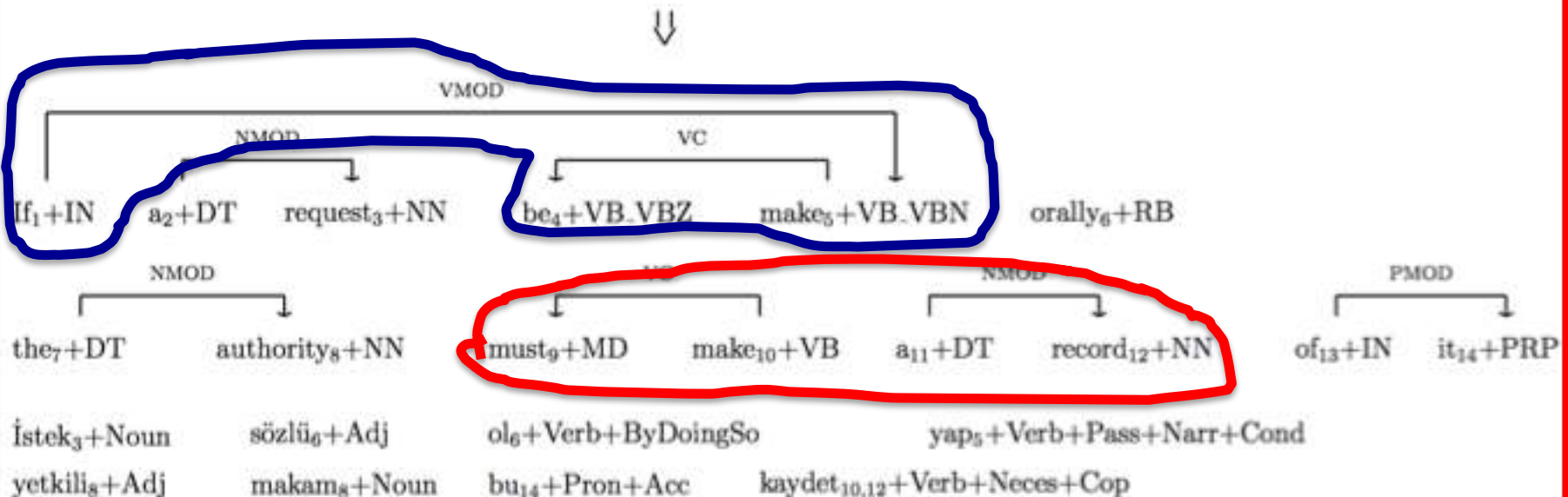


Now align only based on root words – the syntax alignments just follow that

Syntax and Morphology Interaction

If a request is made orally the authority must make a record of it

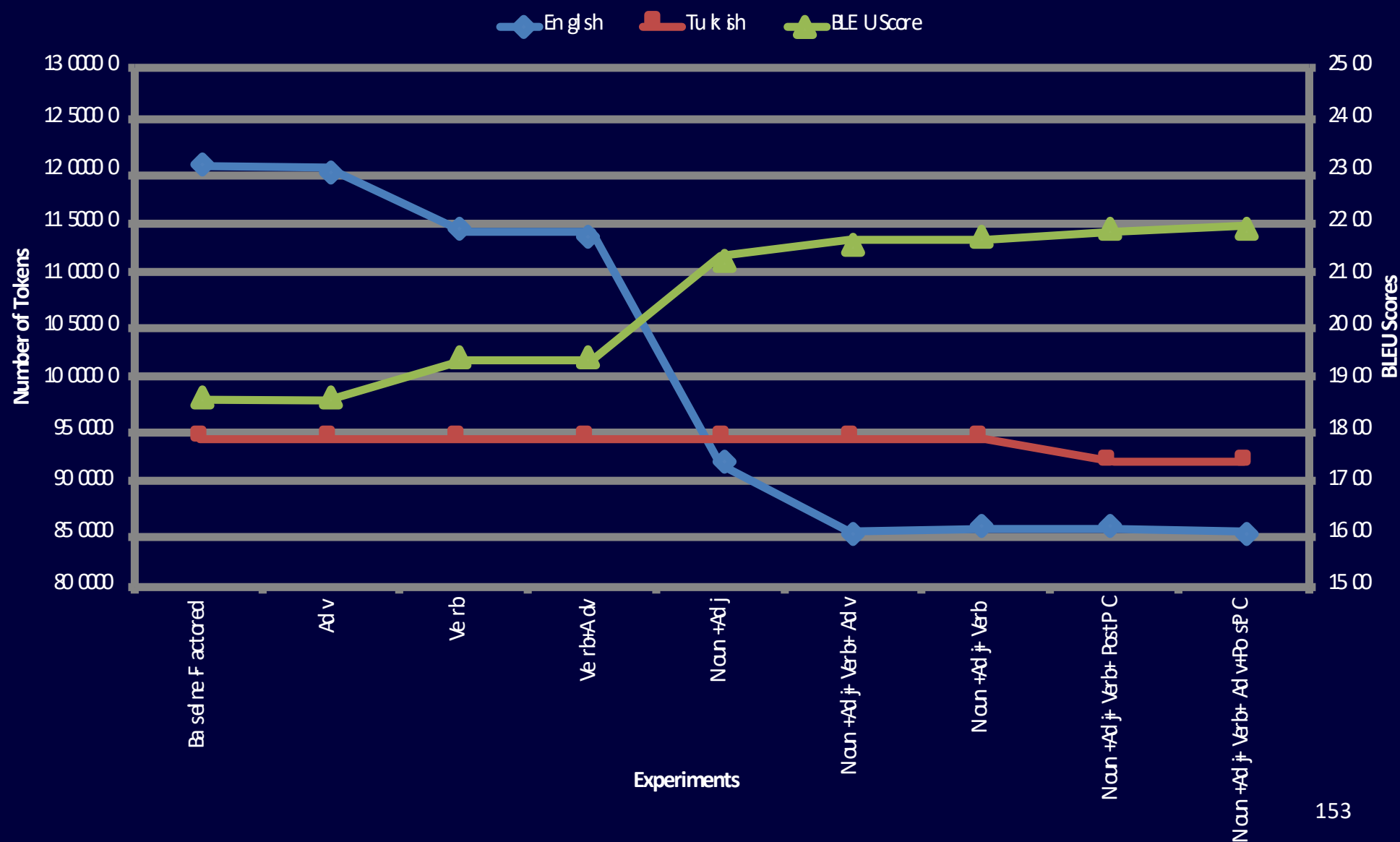
İstek sözlü olarak yapılmışsa yetkili makam bunu kaydetmelidir



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**

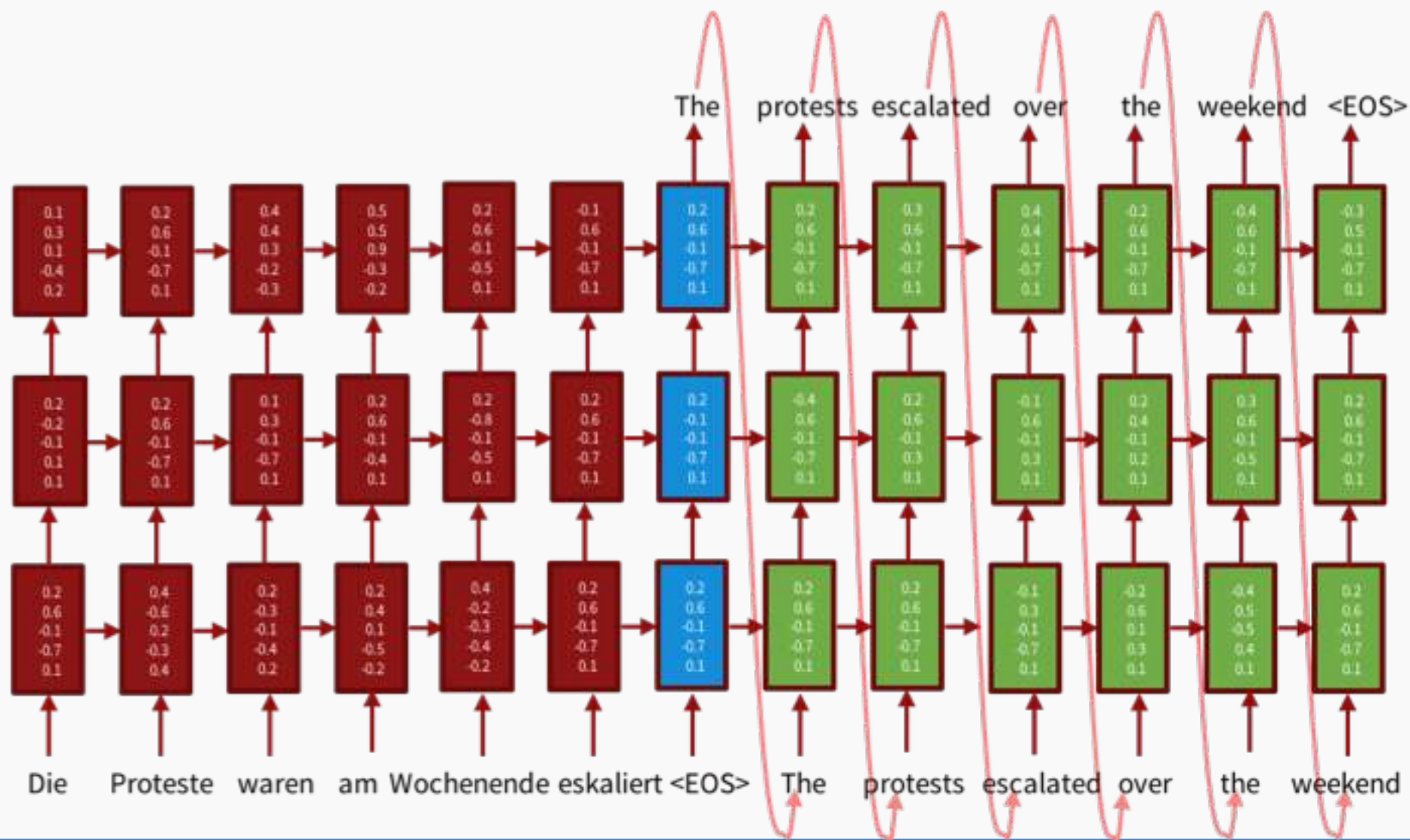
Syntax and Morphology Interaction



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)

Shallow/ Simple

Word-based
only

Phrase table

**Example-
based MT**

Statistical MT

Knowledge
Acquisition
Strategy

Hand-built by
experts

Hand-built by
non-experts

Learn from
annotated data

Learn from un-
annotated data

All manual

Original **direct**
approach

Typical **transfer**
system

Classic
interlingual
system

Syntactic
Constituent
Structure

Semantic
analysis

Interlingua

New Research
Goes Here!

Fully automated

Deep/ Complex

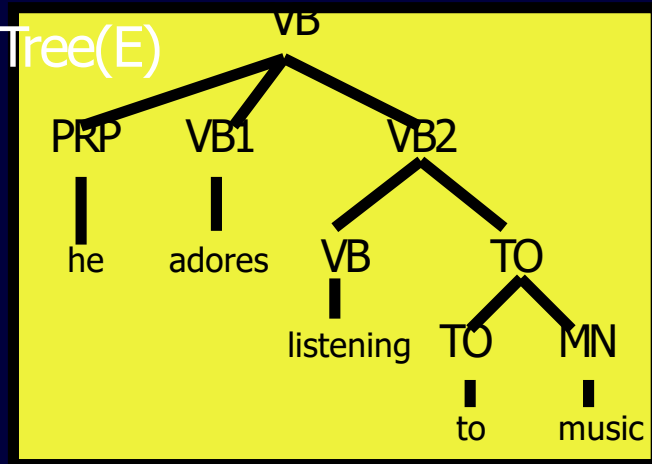
**Knowledge
Representation
Strategy**

Syntax in SMT

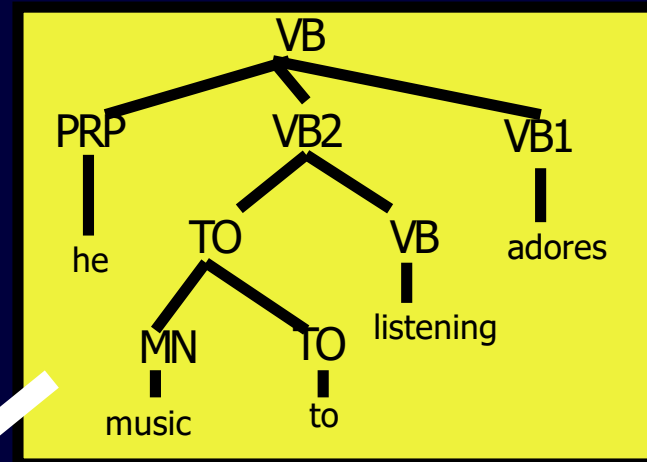
- Early approaches relied on high-performance parsers for one or both languages
 - Good applicability when English is the source language
 - Tree-to-tree or tree-to-string transductions
- Recent approaches induce **synchronous grammars** during training
 - Grammar that describe two languages at the same time
 - $NP \Rightarrow ADJ_{e1} NP_{e2} : NP_{f2} ADJ_{f1}$

Tree-to-String Transformation

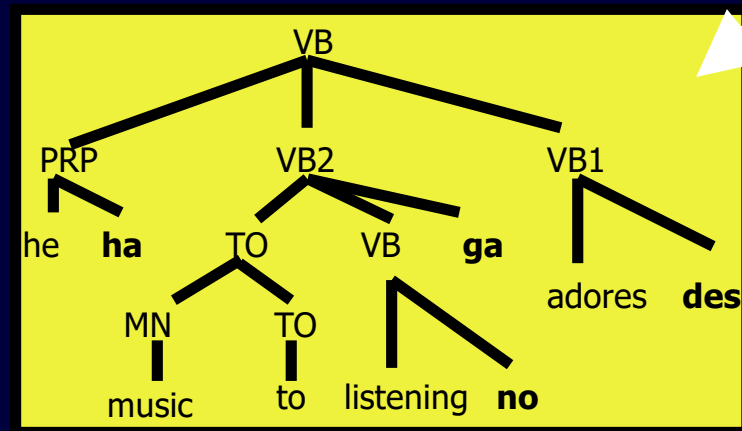
Parse Tree(E)



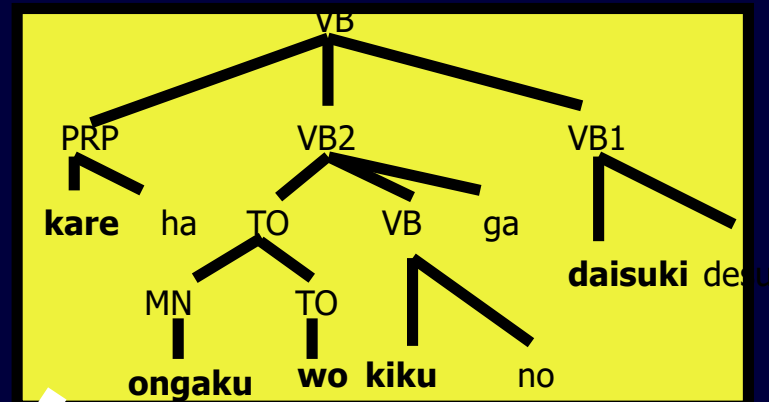
Reorder



Insert



Translate



Take Leaves



Sentence(J)

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)**
 - <https://arxiv.org/pdf/1709.07809.pdf>
- **SMT Workbook (1999)**
 - Kevin Knight
 - Unpublished manuscript at <http://www.isi.edu/~knight/>
- <http://www.statmt.org>
- <http://aclweb.org/anthology-new/>
 - Look for “Workshop on Statistical Machine Translation”