Statistical Machine Translation: Trends & Challenges

2nd International Conference on Arabic Language Resources & Tools
21st April 2009

Prof. Andy Way

NCLT/CNGL,
School of Computing,
Dublin City University,
Dublin 9, Ireland

away@computing.dcu.ie

Dr. Hany Hassan

IBM Cairo HLT Group
IBM Egypt

hanyh@eg.ibm.com

Overview: Part 1 (AW)
14:15 – 16:15

- Why Corpus-Based MT?
- Corpora, and Matters Arising
- Language Modelling
- Translation Models
- Word and Phrase Alignments
- Decoding
- Evaluation

Overview: Part 2 (HH)
16:30 – 18:30

- Factored Models
- Discriminative Training
- Supertag Models of SMT
- Open-Source Tools

Why Corpus-Based MT?

- the (relative) failure of rule-based approaches
- the increasing availability of machine-readable text
- the increase in capability of hardware (CPU, memory, disk space) with decrease in cost

Sine qua non

A prerequisite for Data-Driven MT (and also TM, which is *not* MT, but rather CAT):

- Example-Based MT (EBMT)
- Statistical MT (SMT)
- Hybrid Models which use some probabilistic processing

is a *parallel corpus* (or *bitext*) of aligned sentences.

Corpus-Based MT is here to stay

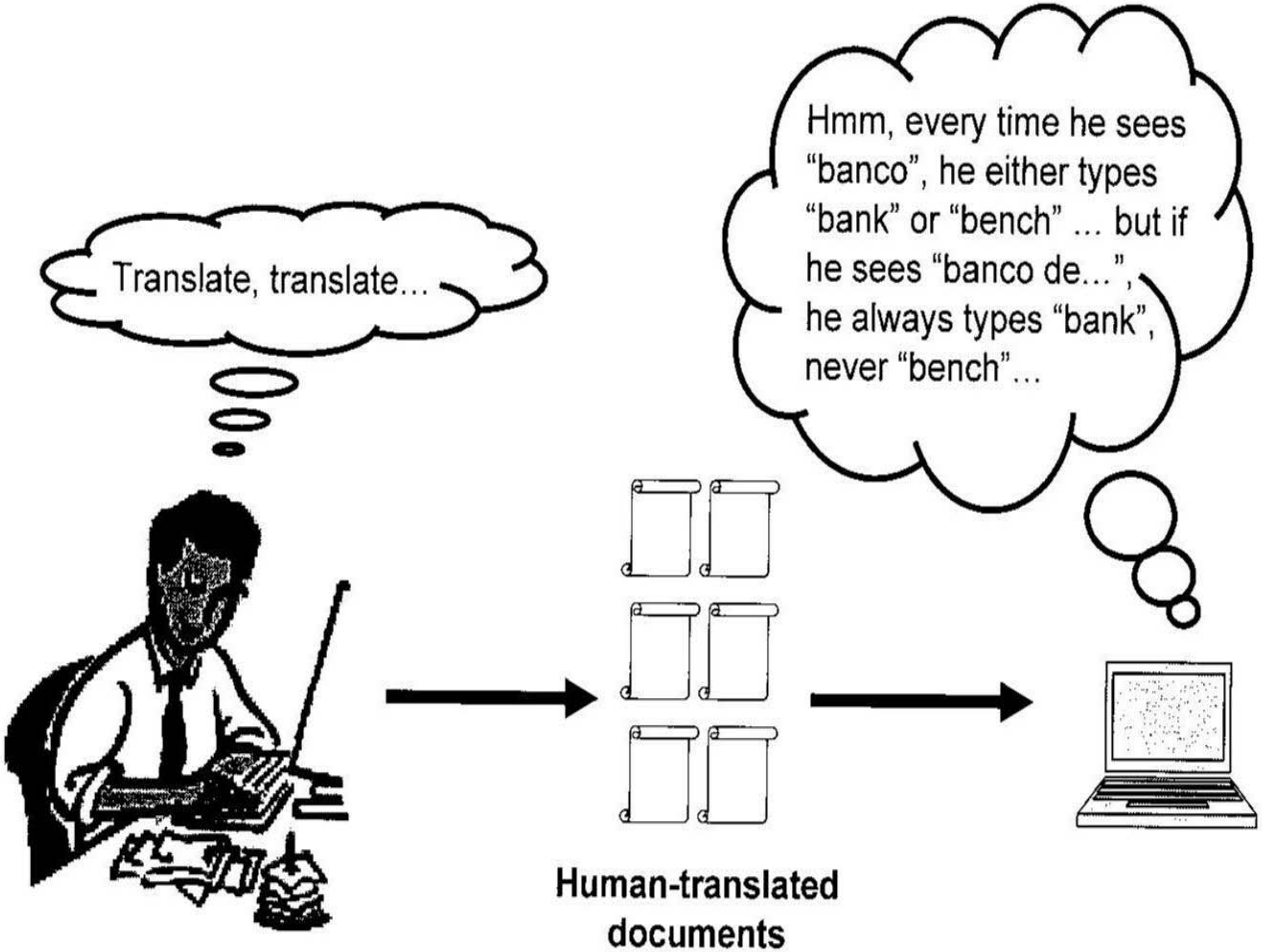
These approaches are now *mainstream*:

- More researchers are developing corpus-based systems;
- 1st company to use SMT now exists: www.languageweaver.com;
- Irish MT company Traslán (www.traslan.ie) uses EBMT;
- In recent large-scale evaluations, corpus-based MT systems come first.

Two caveats:

- Most industrial systems are still rule-based (but cf. Google's online systems now SMT);
- Current mainstream evaluation metrics favour *n*-gram-based systems (i.e. SMT).

Statistical Machine Translation



Thanks to Kevin Knight …

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . cognate?
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order:
{ **jjat, arrat, mat, bat, oloat, at-yurp** }

- There are 6! different orders possible, so 720 different translations.
- Best order (according to placement in TL side of the corpus is as given above):
  - Not just unigrams, but *n*-grams also ...

It’s Really Spanish—English!

Clients do not sell pharmaceuticals in Europe => Clientes 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 .

Some more to try ...

- iat lat pippat eneat hilat oloat at-yurp.
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

Some more to try ...

- iat lat pippat eneat hilat oloat at-yurp.
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

... if you have trouble sleeping at nights!

What did we learn?

- what parallel corpora look like (more on this soon);
- viewing parallel corpora through the ‘eyes’ of a computer;
- how relevant parallel corpora are for MT;
- how to build bilingual dictionaries from parallel corpora;
- how cognate information may be useful in MT;
- how to do word alignment ...

What else do we need to know?

- about word alignment (=dictionary writing) on a larger scale;
- about phrasal alignment, the norm in real translation data;
- about unalignable words;
- the importance of knowing the target language (vs. source) in making fluent translations;
- the importance of short sentence pairs (where alignment possibilities are restricted) in helping disambiguate/align longer sentence pairs;
- about locality in word order shifts;
- how to guess the meanings/translations of unknown words;
- about how much uncertainty the machine faces in working with limited data ...

Can such methods be scaled to ‘real’ MT?

- Availability of monolingual and bilingual corpora?
- Possibility of sentence-aligning bilingual corpora?
- Can we write an algorithm to extract the translation dictionary?
- Can we write an algorithm to extract the monolingual word pair counts?
- Can we write an algorithm to generate translations using our translation dictionary and word pair counts?

Can such methods be scaled to ‘real’ MT?

- Availability of monolingual and bilingual corpora?
- Possibility of sentence-aligning bilingual corpora?
- Can we write an algorithm to extract the translation dictionary?
- Can we write an algorithm to extract the monolingual word pair counts?
- Can we write an algorithm to generate translations using our translation dictionary and word pair counts?

- WILL THE TRANSLATIONS PRODUCED BE ANY GOOD?

Parallel Corpora

- Hugely important ... but not available in a wide range of language pairs:
  - Chinese—English: Hong Kong data
  - French—English: Canadian Hansards
  - Older EU pairs: Europarl [Koehn 04]
  - Newer EU pairs: JRC-Acquis Communautaire
  - Arabic—English: LDC Data
  - NIST, IWSLT, WMT, TC-STAR Evaluations

Good Quality Language & Translation Models

- Any statistical approach to MT requires the availability of aligned bilingual corpora which are:
  - large;
  - good-quality;
  - representative.

Corpus 1

Mary and John have two children.
The children that Mary and John have are aged 3 and 4.
John has blue eyes.

Question 1: what's P(have) vs. P(has) in a corpus?

Question 2: what's P(have |John) vs. P(has | John) in a corpus?

Question 3: what's P(have) vs. P(has) in *this* corpus? What's their *relative* probability?

Question 4: what's P(have | John) vs. P(has | John) in *this* corpus?

Corpus 2

Am I right, or am I wrong?
Peter and I are seldom wrong.
I am sometimes right.
Sam and I are often mistaken.

Question 5: What two generalisations would a probabilistic language model (based on *bigrams*, say) infer from this data, which are not true of English as a whole? Are there any other generalisations that could be inferred?

Question 6: Try to think of some trigrams (and 4-grams, if you can) that cannot be 'discovered' by a bigram model? What you're looking for here is a phrase where the third (or subsequent) word depends on the first word, which in a bigram model is 'too far away' ...

Some Observations

- Note that all the sentences in these corpora are well-formed.
- If, on the other hand, the corpus contains ill-formed input, then that too will skew our probability models ...

... and our translations will be affected!

Corpus 1 Revisited

- Using Google on 10th February 2003, I got:
  - # ‘have’ = 380,000,000
  - # ‘has’ = 244,000,000
  - # ‘John has’ = 227,000
  - # ‘John have’ = 25,700
- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

Corpus 2 Revisited

- Using Google on 10th February 2003, I got:

    - # ‘am I’ = 3,690,000
    - # ‘I am’ = 8,060,000
    - # ‘I are’ = 1,230,000

- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

Bilingual Corpora

All this applies to bitexts too!

Q: of what English word are these possible French translations (from the *Canadian Hansards*, note)?

Q: what's ???

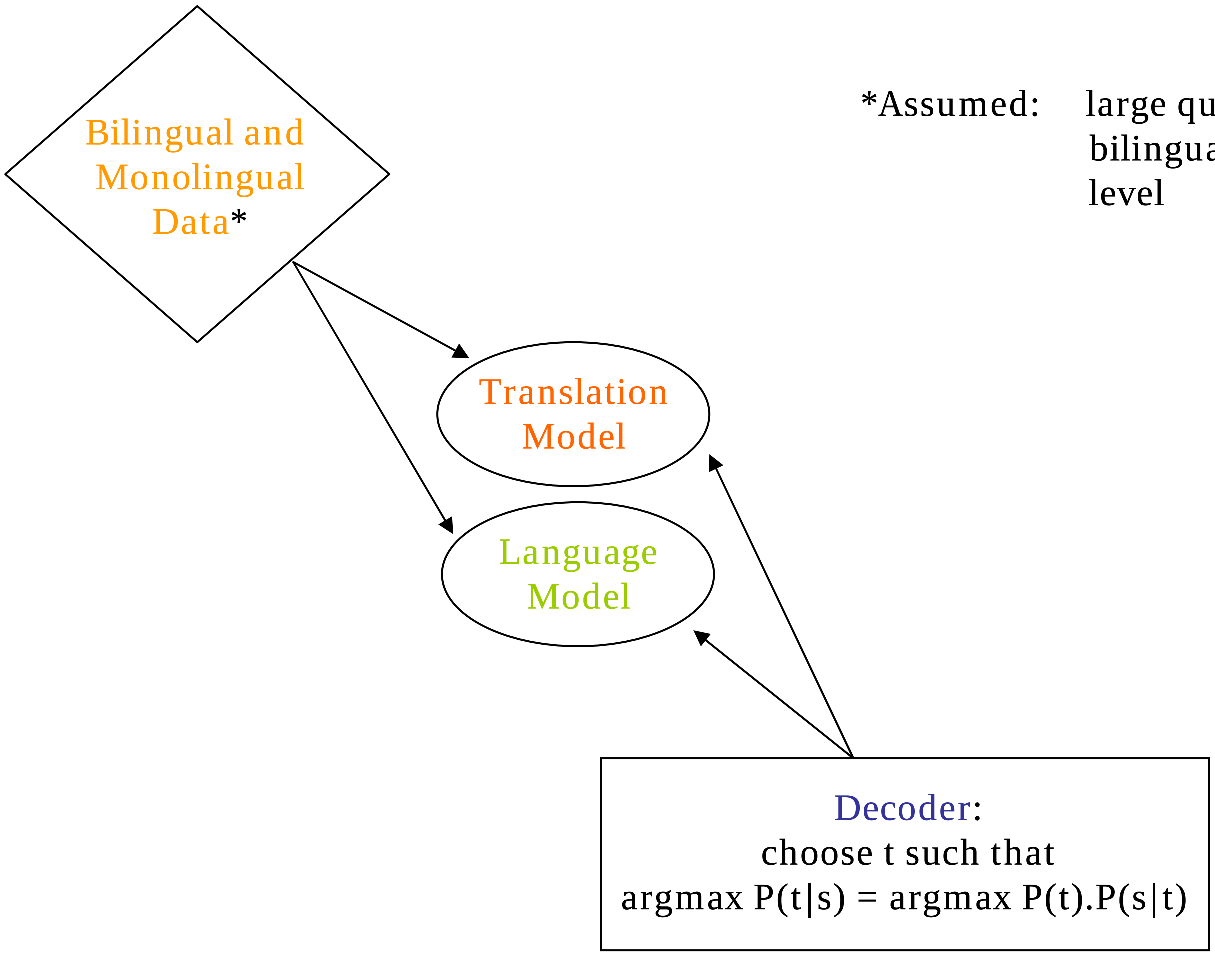
French	**Probability**
???	.808
entendre	.079
entendu	.026
entends	.024
entendons	.013

Caveat interpres!

- Beware of sparse data!
- Beware of unrepresentative corpora!
- Beware of poor quality language!

If the corpora are small, or of poor quality, or are unrepresentative, then our statistical language models will be poor, so any results we achieve will be poor.

Statistical Machine Translation (SMT)



*Assumed: large quantities of high-quality bilingual data aligned at sentence level

Thanks to Mary Hearne for some of these slides

Basic Probability

Consider that any source sentence *s* may translate into any target sentence *t*. It's just that some translations are more likely than others. How do we formalise "more likely"?

<u>P(s) : *a priori* probability</u>

The chance/likelihood/probability that s happens.
For example, if s is the English string "I like spiders", then P(s) is the likelihood that some person at some time will utter the sentence "I like spiders" as opposed to some other sentence.

<u>P(t|s) : *conditional* probability</u>

The chance/likelihood/probability that t happens *given that s has happened*.
If s is again the English string "I like spiders" and t is the French string "Je m'appelle Andy" then P(t|s) is the probability that, upon seeing sentence s, a translator will produce t.

Basic Probability

<u>$P(s,t)$: *joint* probability</u>

The chance/likelihood/probability that s and t both happen.

If s and t don't influence each other then we can say:

$$P(s,t) = P(s) * P(t)$$

However, if s and t are mutual translations then this doesn't hold, so we say:

$$P(s,t) = P(s) * P(t|s)$$

In English: the chances of s and t both happening is equal to the chances of s happening anyway (independently of t) multiplied by the chances of t happening given that we've already seen s.

All probabilities are between 0 and 1 inclusive.
A probability of 0.5 means "there's half a chance".

What's the probability of throwing at least 7 using two dice?

What's the probability of throwing at least 7 given that you've already thrown 6 on the first dice?

Sums and Products

To represent the addition of integers from 1 to n:

$$\sum_{i=1}^{n} i \qquad (= 1 + 2 + 3 + 4 + \ldots + n)$$

If everything being summed over is multiplied by a factor then this can be taken outside:

$$\sum_{i=1}^{n} i * k \quad = 1k + 2k + 3k + 4k + \ldots + nk = \quad k \sum_{i=1}^{n} i$$

To represent the multiplication of integers from 1 to n:

$$\prod_{i=1}^{n} i \qquad (= 1 * 2 * 3 * 4 * \ldots * n)$$

Language Modelling

- A language model assigns a probability to *every* string in that language. We've done some of this already with our toy corpora.

- In practice, we gather a huge database of utterances and then calculate the relative frequencies of each.

We could use the Web …

I like spiders.
I hate spiders.
I hate spiders that are poisonous.

- We just count how many of each there are and give their relative frequency …
- Problem 1: many (nearly all) strings will receive *no* probability as we haven’t seen them …
- Problem 2: all unseen good and bad strings are deemed equally unlikely …
- Solution? How do we know if a new utterance is valid or not? By breaking it down into substrings (‘constituents’?)

Language Modelling

- We've already dealt with substrings, or *n*-grams.

Hypothesis:

If a string has lots of reasonable/plausible/likely *n*-grams then it might be a reasonable sentence.

How do we measure plausibility, or 'likelihood'?

n-grams

Suppose we have the phrase “x y” (i.e. word “x” followed by word “y”).

P(y|x) is the probability that word y follows word x

A commonly-used n-gram estimator:

$$P(y|x) = \frac{\text{number-of-occurrences (“x y”)}}{\text{number-of-occurrences (“x”)}}$$

Bigrams

Similarly, suppose we have the phrase “x y z”.

P(z|x y) is the probability that word z follows words x and y

$$P(z|x\ y) = \frac{\text{number-of-occurrences (“x y z”)}}{\text{number-of-occurrences (“x y”)}}$$

Trigrams

Language Modelling

N-gram language models can assign non-zero probabilities to sentences they have never seen before:

P("I don't like spiders that are poisonous") =
P("I don't like") *
P("don't like spiders") *
P("like spiders that") *
P("spiders that are") *
P("that are poisonous")
> 0 ?

Trigrams …

Bigrams …

P("I don't like spiders that are poisonous") =
P("I don't") *
P("don't like") *
P("like spiders") *
P("spiders that") *
P("that are") *
P("are poisonous")
> 0 ?

Or even Unigrams, or more likely <u>some weighted combination of all these</u>

Language Modelling

Building n-gram models for larger values of n is often impractical due to the large numbers of parameters (or n-gram probabilities) which have to be estimated.

Suppose, for example, that we have a corpus containing 20,000 word types:

Model	Number of parameters
bigram	Approx. $20{,}000^2$ = 400 million
trigram	Approx. $20{,}000^3$ = 8 trillion
4-gram	Approx. $20{,}000^4 = 1.6 \times 10^{17}$

Ways of reducing the number of parameters:

- reduce the value of n
- stem the words (removing inflectional endings)
- group words into semantic classes
- condition on, for example, previous word + predicate

However, n-gram models are the simplest to work with.

Language Modelling

Building n-gram models for larger values of n is often impractical due to the large numbers of parameters (or n-gram probabilities) which have to be estimated.

Suppose, for example, that we have a corpus containing 20,000 word types:

Model	Number of parameters
bigram	Approx. $20{,}000^2$ = 400 million
trigram	Approx. $20{,}000^3$ = 8 trillion
4-gram	Approx. $20{,}000^4$ = 1.6×10^{17}

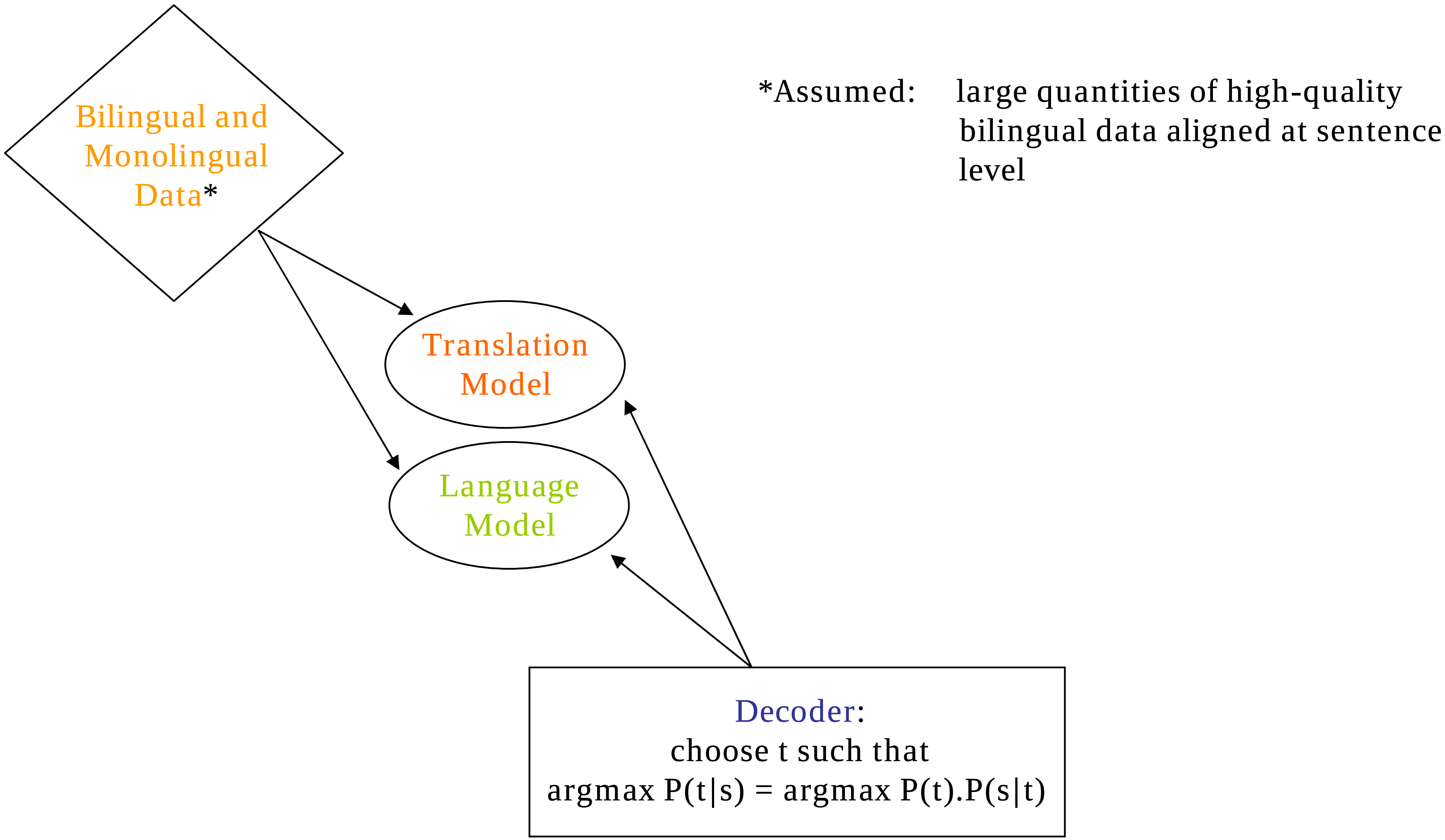
Comparison (thanks to Chris Callison-Burch): the no. of milliseconds until the sun becomes a red giant and engulfs the Earth ≈ 1.6×10^{20}

Ways of reducing the number of parameters:

- reduce the value of n
- stem the words (removing inflectional endings)
- group words into semantic classes
- condition on, for example, previous word + predicate

However, n-gram models are the simplest to work with.

Statistical Machine Translation (SMT)



The Translation Model

the language model

$$\text{argmax } P(t|s) = \text{argmax } P(t) \,.\, P(s|t)$$

the translation model

<u>At its simplest:</u>

the translation model needs to be able to take a bag of L*x* words and a bag of L*y* words and establish how likely it is that they correspond.

<u>Or, in other words:</u>

the translation model needs to be able to turn a bag of L*x* words into a bag of L*y* words and assign a score P(t|s) to the bag pair.

The Translation Model

the language model

SMT: $\operatorname{argmax} P(e|f) = \operatorname{argmax} P(e) \cdot P(f|e)$

the translation model

<u>Remember:</u>

If we carry out, for example, French-to-English translation, then we will have:

- an English Language Model, and
- an English-to-French Translation Model.

When we see a French string f, we want to reason backwards ... What English string e is:

- likely to be uttered?
- likely to then translate to f?

We are looking for the English string e that maximises

$$P(e) * P(f|e).$$

The Translation Model

<u>Word re-ordering in translation:</u>

The language model establishes the probabilities of the possible orderings of a given bag of words, e.g.

{*have,programming,a,seen,never,I,language,better*}.

Effectively, the language model worries about word order, so that the translation model doesn't have to...
But what about a bag of words such as

{*loves,John,Mary*}?

Maybe the translation model *does* need to know a little about word order, after all...

The Translation Model

IBM Model 3

P. Brown et al. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* **19**(2):263—311.

Translation as string re-writing:

John did not slap the green witch

FERTILITY

John not slap slap slap the green witch

TRANSLATION

John no daba una botefada la verde bruja

INSERTION

John no daba una botefada a la verde bruja

DISTORTION

John no daba una botefada a la bruja verde

The Translation Model

n: fertility parameters, e.g.

n(1|house) = ?
n(2|house) = ?
n(3|house) = ?
...

i.e. what is the probability that "house" will produce exactly 1/2/3 French words whenever "house" appears?

t: word-translation parameters, e.g.

t(maison|house) = ?
t(domicile|house) = ?
t(amelioration|house) = ?
...

i.e. what is the probability that "house" will produce the French word maison/domicile/amelioration whenever "house" appears?

The Translation Model

d: distortion parameters, e.g.

$d(2|2) = ?$
$d(3|2) = ?$
$d(5|2) = ?$
...

i.e. what is the probability that the English word in position 2 of the English sentence will generate a French word that winds up in position 2/3/5... of a French translation?

p: We also have word-translation parameters corresponding to insertions:

p(à|NULL) = ?
p(de|NULL) = ?
p(pour|NULL) = ?
...

i.e. what is the probability that the French word à/de/pour is <u>inserted</u> into the French string?

Summary of Translation Model Parameters

FERTILITY	n	table plotting source words against fertilities
TRANSLATION	t	table plotting source words against target words
INSERTION	p	single number indicating the probability of insertion
DISTORTION	d	table plotting source string positions against target string positions

Summary of Translation Model Parameters

FERTILITY	n	table plotting source words against fertilities
TRANSLATION	t	table plotting source words against target words
INSERTION	p	single number indicating the probability of insertion
DISTORTION	d	table plotting source string positions against target string positions

How can we automatically obtain parameter values t, n, d and p from data?

Via the EM Algorithm!

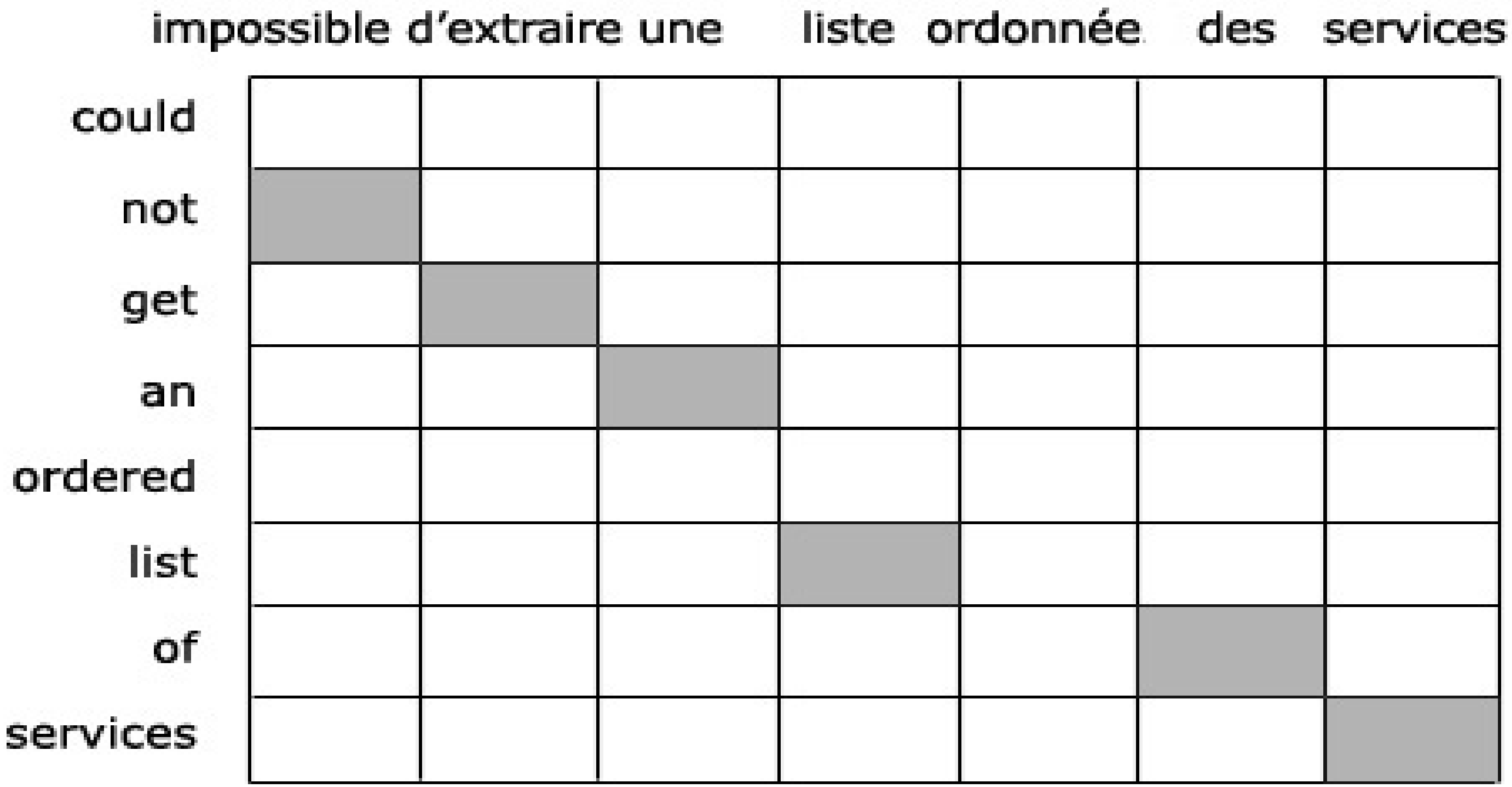
Phrasal Alignments in SMT

- Everything we've looked at so far assumes a set of word alignments.
- As speakers of foreign languages, we know that words don't map one-to-one.
- It'd be better if we could map 'phrases', or sequences of words, and if need be probabilistically reorder them in translation ...

Advantages of Phrasal Alignments

- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguation:
  - Interest in → …
  - Interest rate → …
- The more data, the longer the learned phrases (whole sentences, sometimes …)

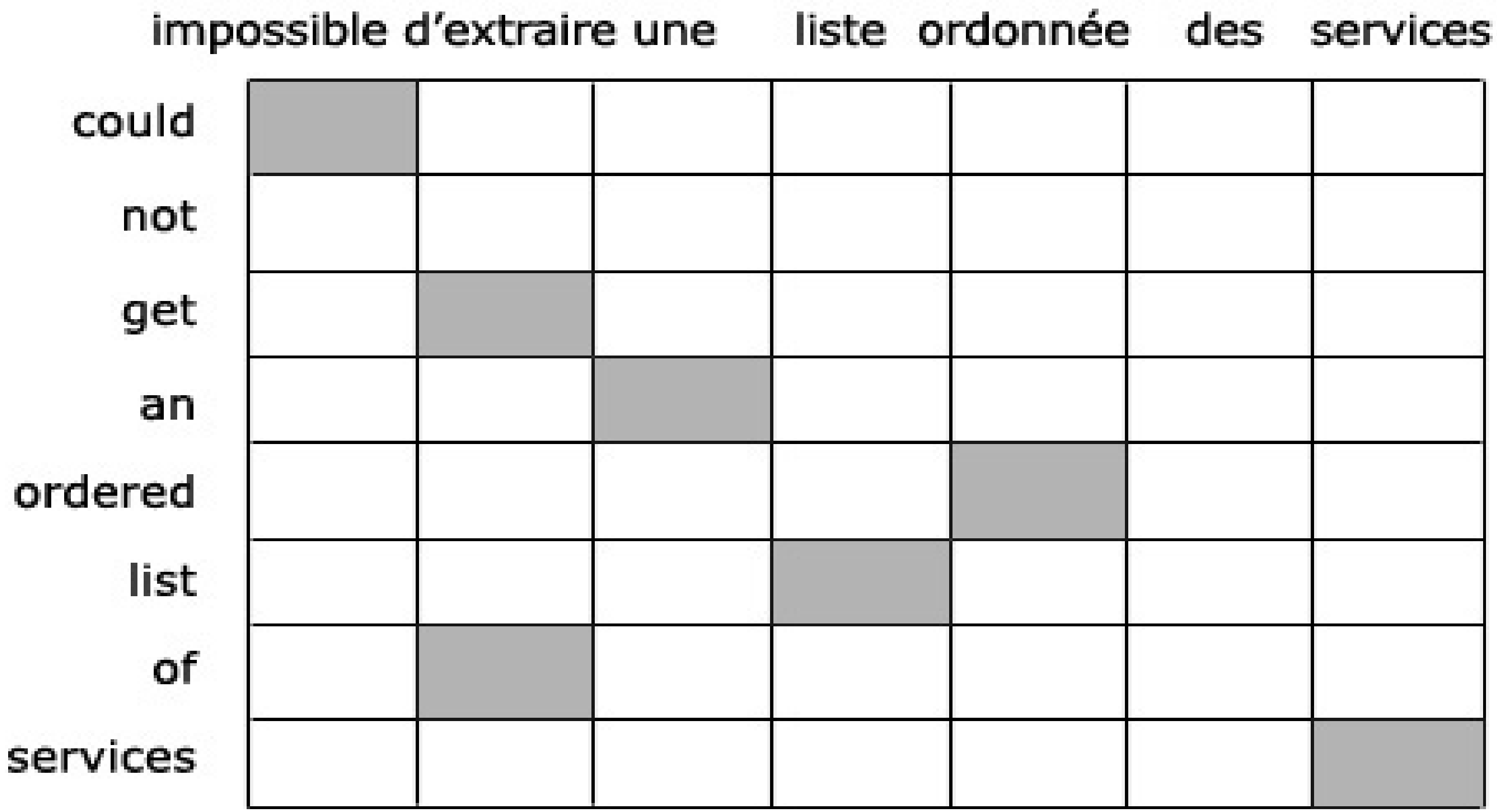
Learning Phrasal Alignments



Here's a set of English→French Word Alignments

Thanks to Declan Groves for these …

Learning Phrasal Alignments

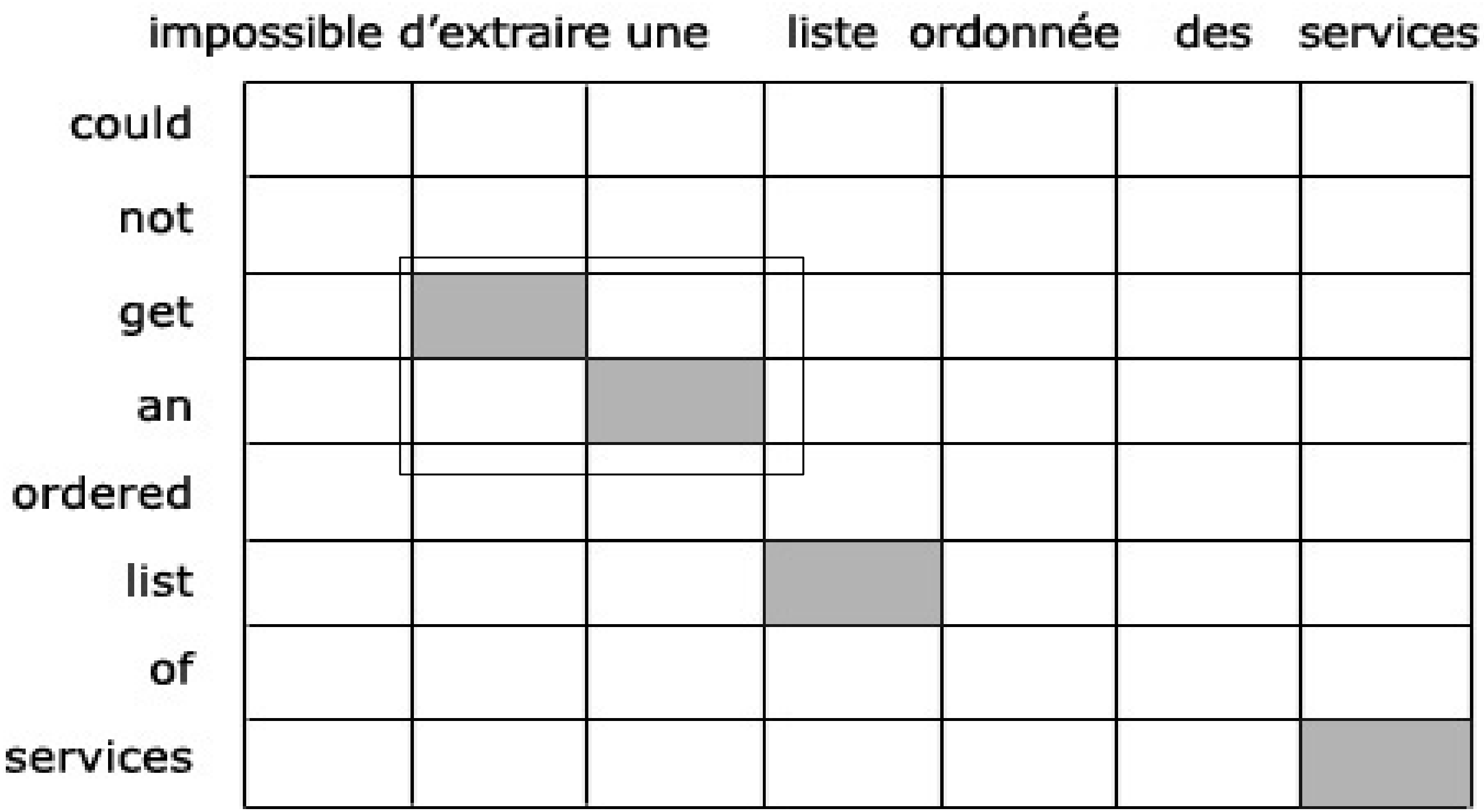


Here's a set of French→English Word Alignments

Learning Phrasal Alignments

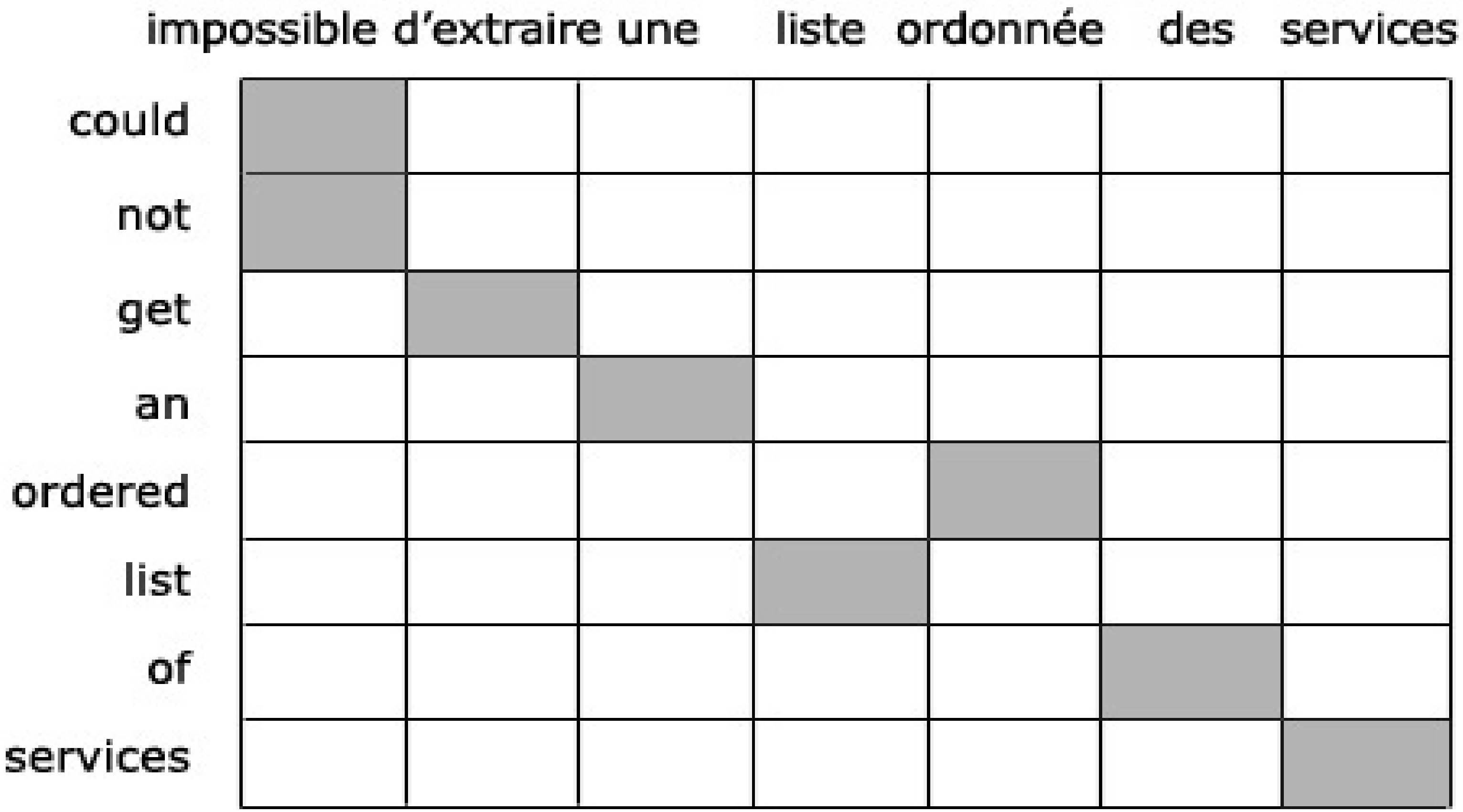
We can take the Intersection of both sets of Word Alignments

Learning Phrasal Alignments



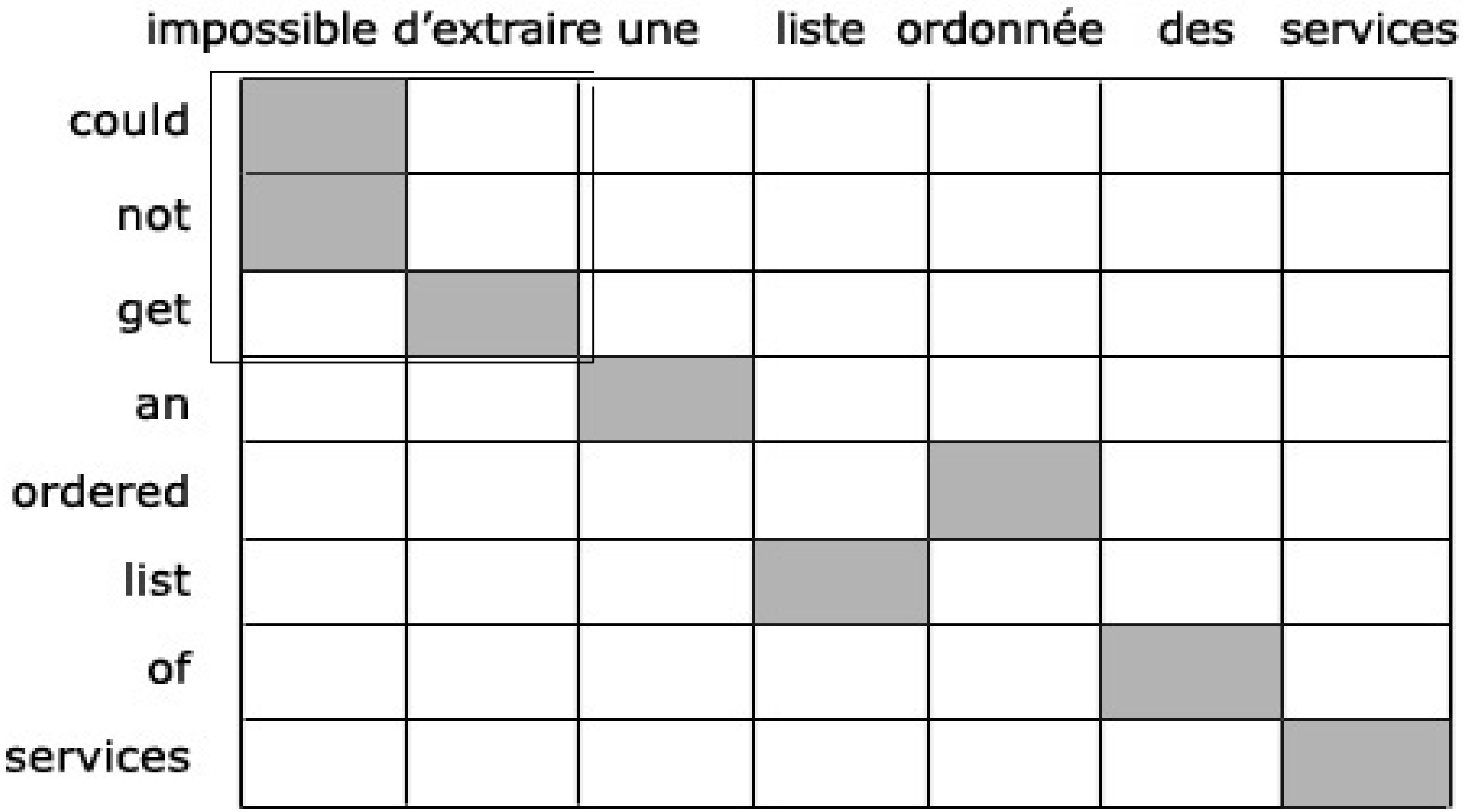
Taking contiguous blocks from the Intersection
gives sets of highly confident phrasal Alignments

Learning Phrasal Alignments



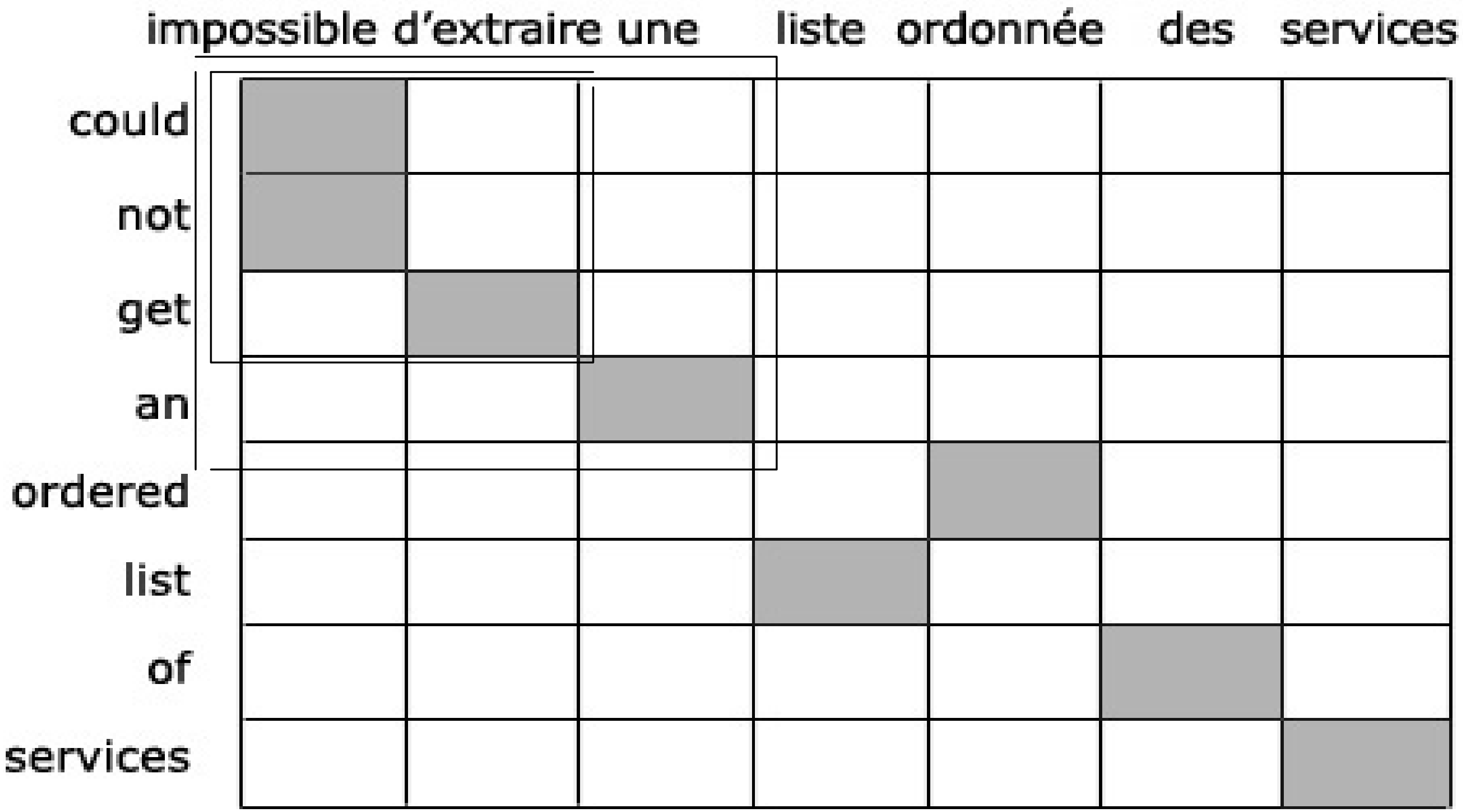
And back off to the Union of both sets of Word Alignments

Learning Phrasal Alignments



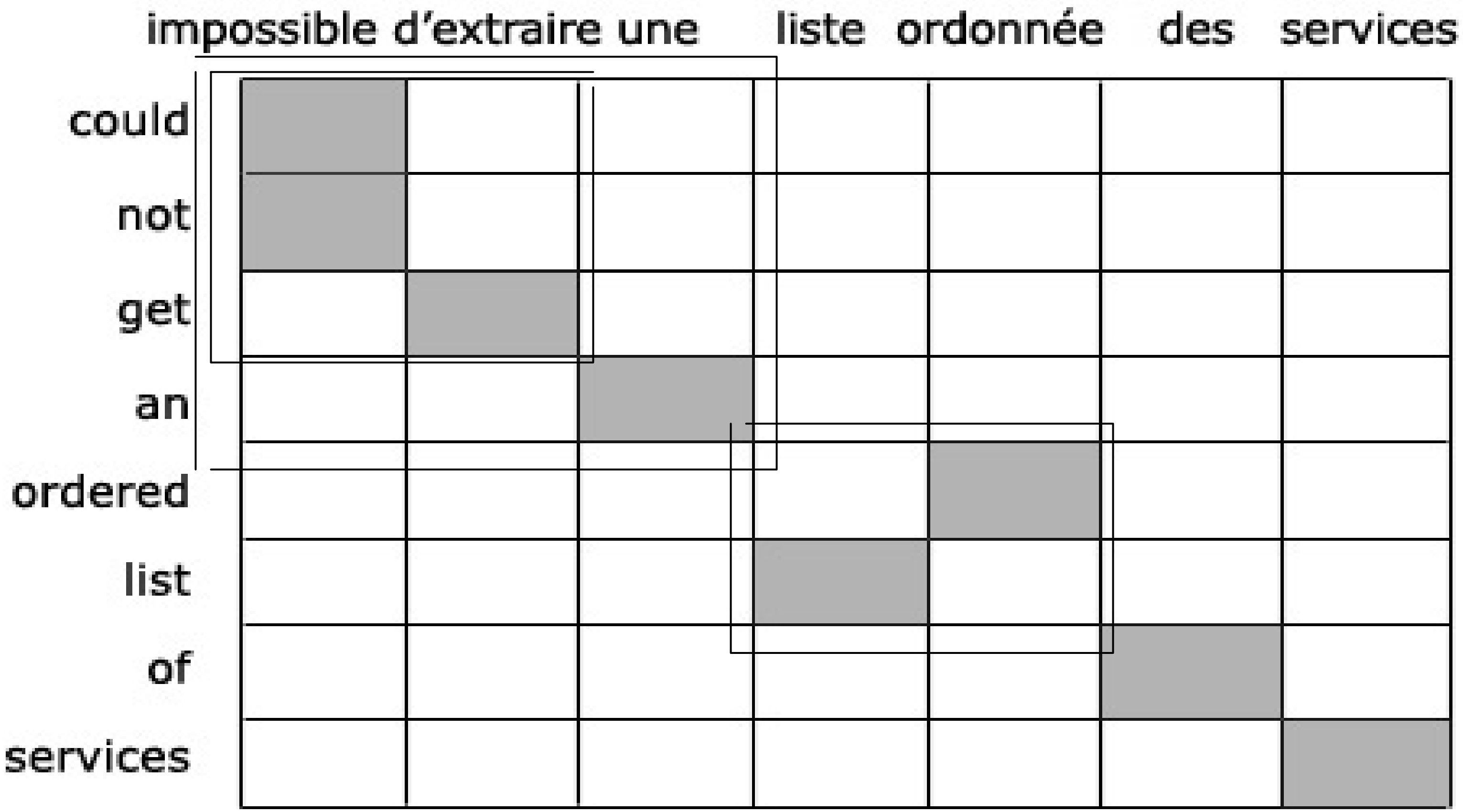
We can also group together contiguous blocks from the Union
to give us (less confident) sets of phrasal alignments

Learning Phrasal Alignments



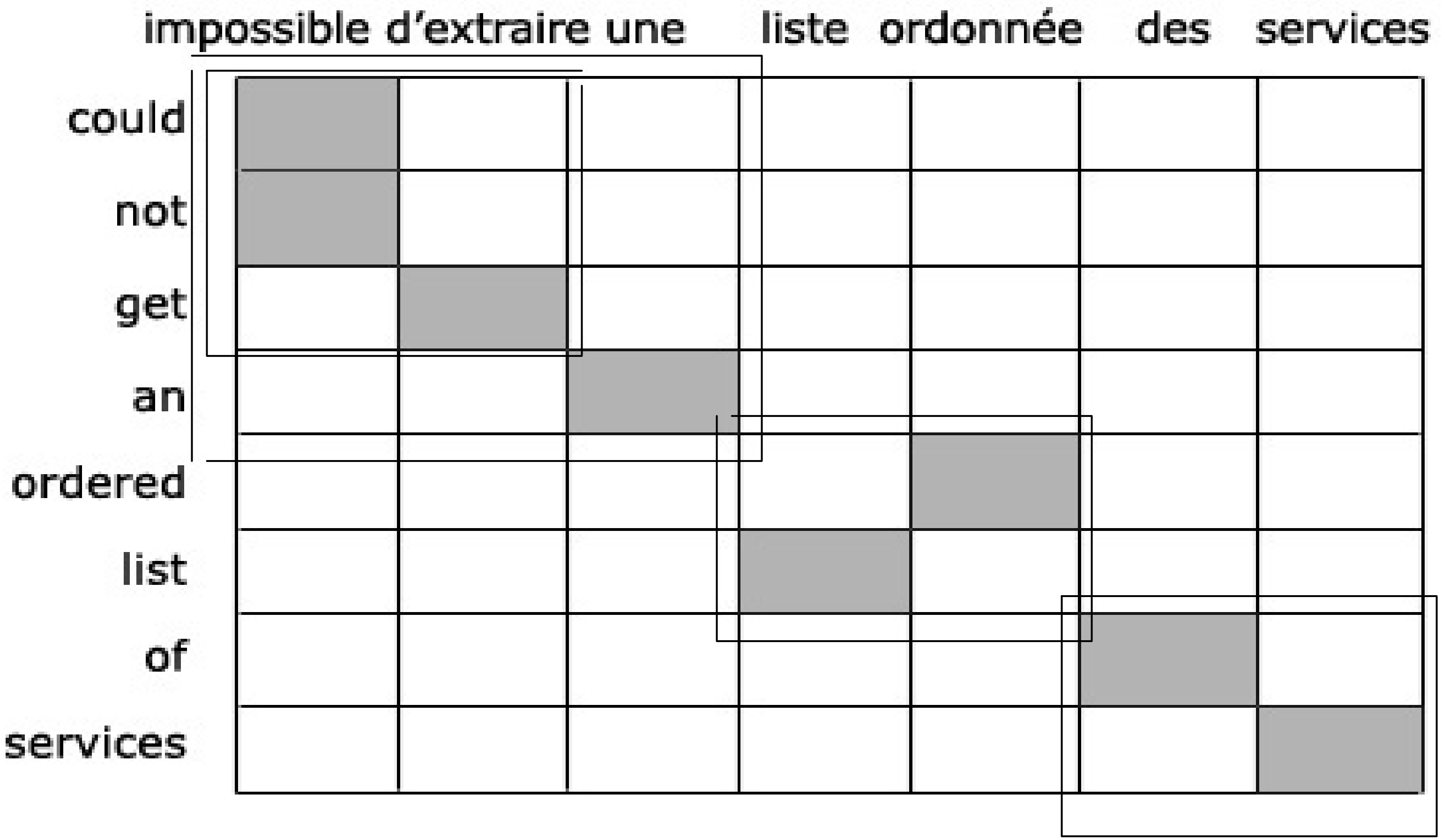
We can also group together contiguous blocks from the Union
to give us (less confident) sets of phrasal alignments

Learning Phrasal Alignments



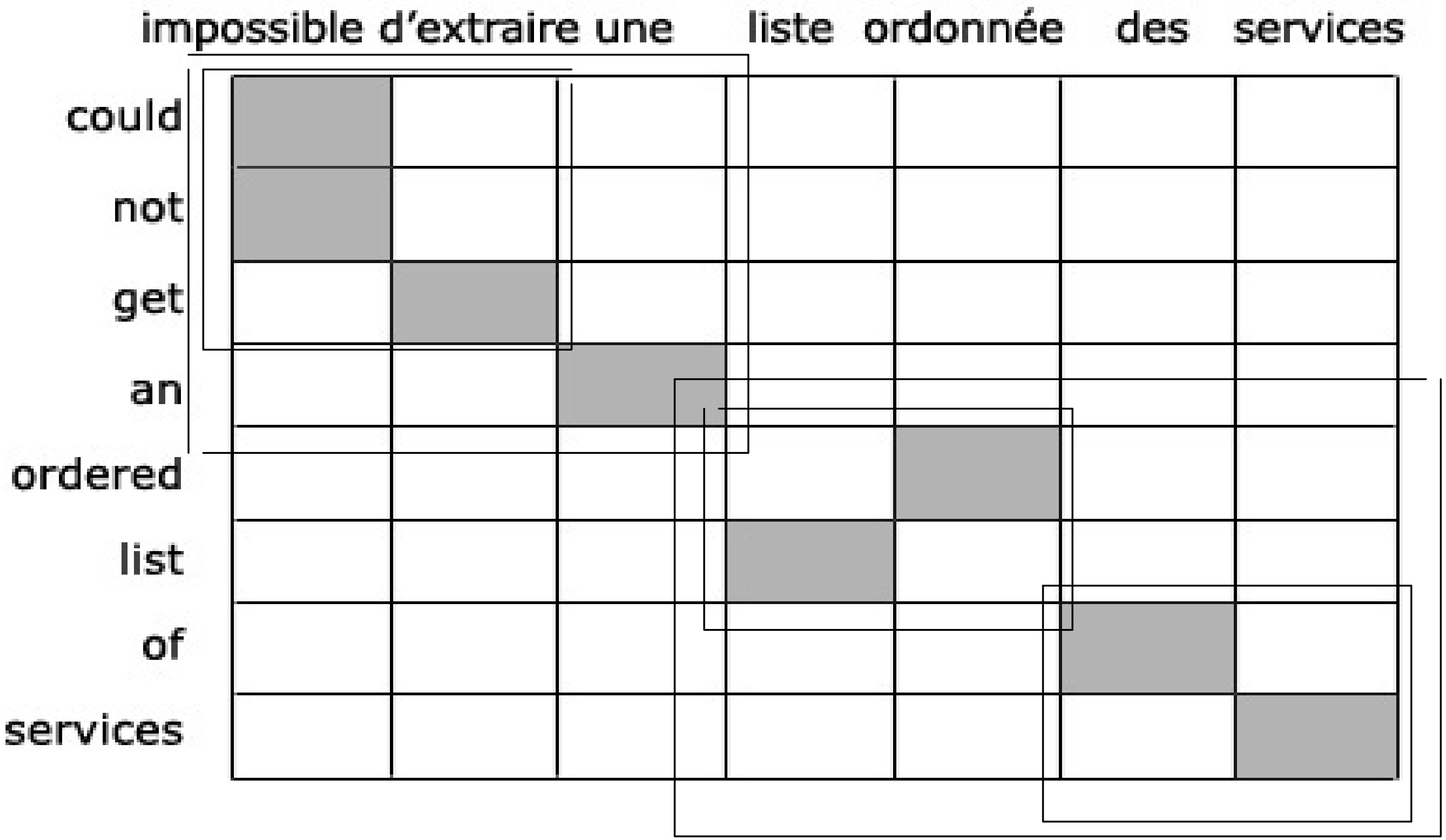
We can also group together contiguous blocks from the Union
to give us (less confident) sets of phrasal alignments

Learning Phrasal Alignments



We can also group together contiguous blocks from the Union
to give us (less confident) sets of phrasal alignments

Learning Phrasal Alignments



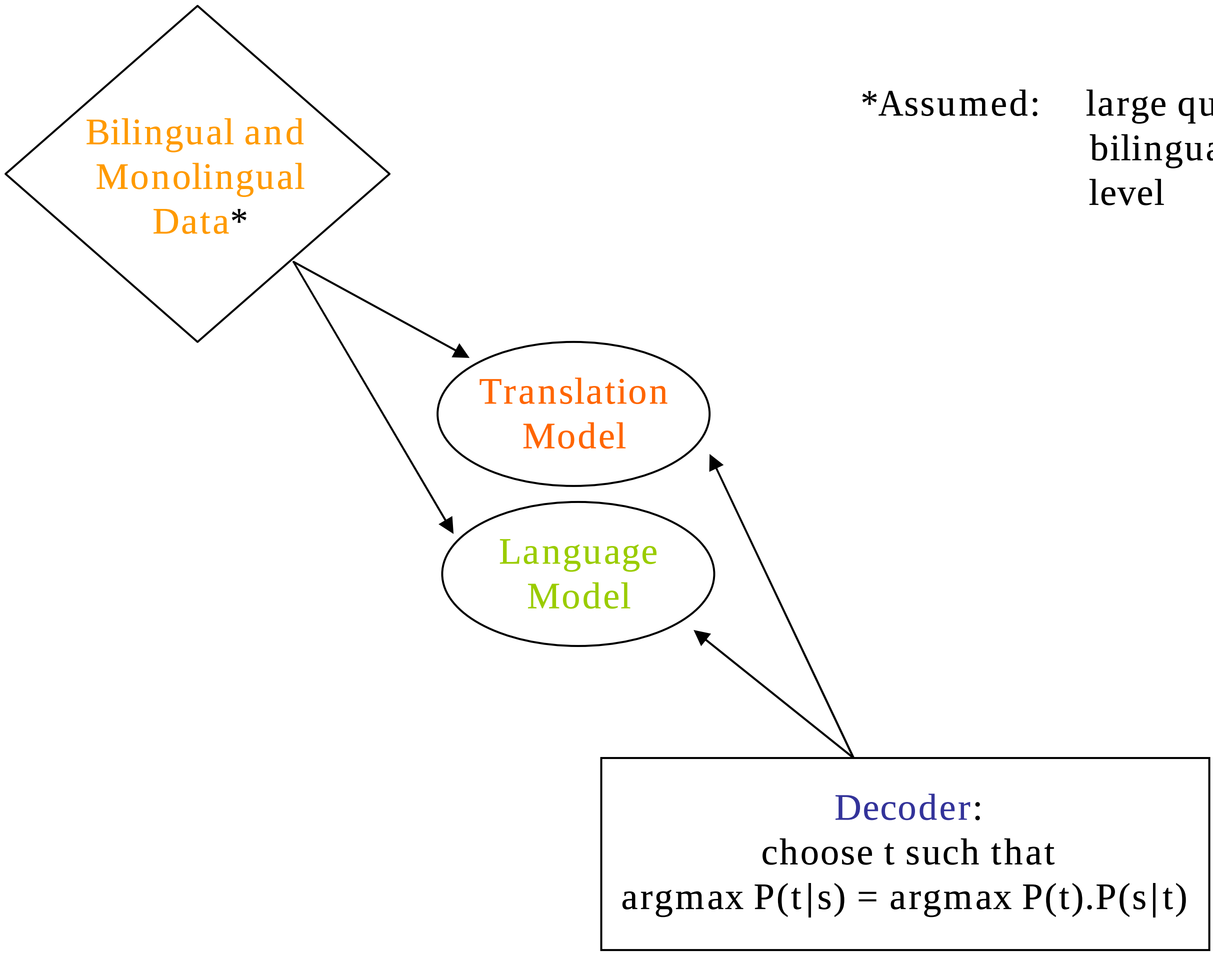
We can also group together contiguous blocks from the Union
to give us (less confident) sets of phrasal alignments

Learning Phrasal Alignments

- We can learn as many phrase-to-phrase alignments as are consistent with the word alignments
- EM training and relative frequency can give us our phrase-pair probabilities
- One alternative is the joint phrase model [Marcu & Wang 02; Birch et al., 06]

Statistical Machine Translation (SMT)

*Assumed: large quantities of high-quality bilingual data aligned at sentence level



Decoding

- given input string s, choose the target string t that maximises P(t|s)

$$\text{argmax } P(t|s) = \text{argmax } (\, P(t) * P(s|t) \,)$$

Language Model

Translation Model

Decoding

- Monotonic version:
  - Substitute phrase by phrase, left to right
  - Word order can change within phrases, but phrases themselves don’t change order
  - Allows a dynamic programming solution (beam search)
  - Monotonic assumption not as damaging as you’d think (for Arabic/Chinese—English, about 3—4 BLEU points)
- Non-monotonic version:
  - Explore reordering of phrases themselves

Decoding Process

Maria no dio una botefada a la bruja verde

- Build translation left to right
  - Select foreign words to be translated

Thanks to Phillip Koehn for these …

Decoding Process

Maria no dio una botefada a la bruja verde

$\downarrow$

Mary

- Build translation left to right
  - Select foreign words to be translated
  - Find English phrase translation
  - Add English phrase to end of partial translation

Decoding Process

Maria no dio una botefada a la bruja verde

Mary

- Build translation left to right
  - Select foreign words to be translated
  - Find English phrase translation
  - Add English phrase to end of partial translation
  - Mark words as translated

Decoding Process

Maria no dio una botefada a la bruja verde

Mary did not

- One to many translation

Decoding Process

Maria no dio una botefada a la bruja verde

Mary did not slap

•Many to one translation

Decoding Process

Maria no dio una botefada a la bruja verde

Mary did not slap the

•Many to one translation

Decoding Process

Maria no dio una botefada a la bruja verde

Mary did not slap the green

•Reordering

Decoding Process

Maria no dio una botefada a la bruja verde

Mary did not slap the green witch

- Translation finished

Translation Options

Maria no dio una botefada a la bruja verde

Mary not give a slap to the witch green

did not a slap by green witch

no slap to the

slap the witch

- Look up possible phrase translations
  - Many different ways to segment words into phrases
  - Many different ways to translate each phrase

Hypothesis Expansion

Maria no dio una botefada a la bruja verde

Mary not give a slap to the witch green

did not a slap by green witch

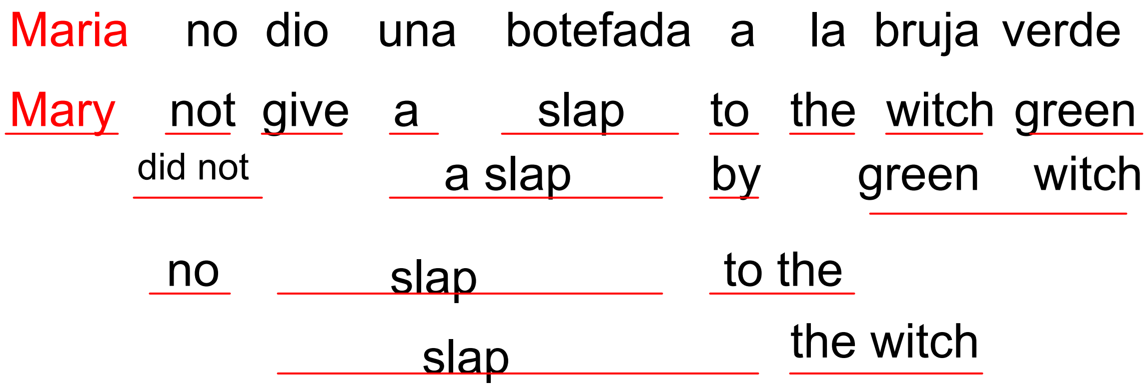
no slap to the

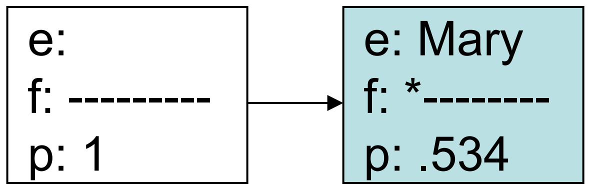
slap the witch

- Start with empty hypothesis
  - e: no English words
  - f: no foreign words covered
  - p: probability 1

e:
f: -----
p: 1

Hypothesis Expansion





- Pick translation option
- Create hypothesis
    - e: add English phrase ‘Mary’
    - f: first foreign word covered
    - p: probability .534

Hypothesis Expansion

Maria no dio una botefada a la bruja verde

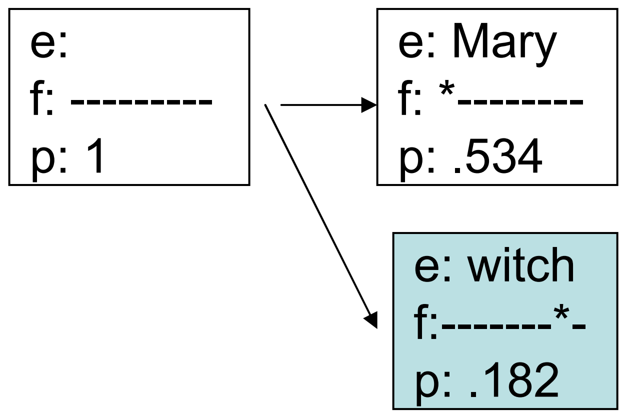
Mary not give a slap to the witch green

did not a slap by green witch

no slap to the

slap the witch

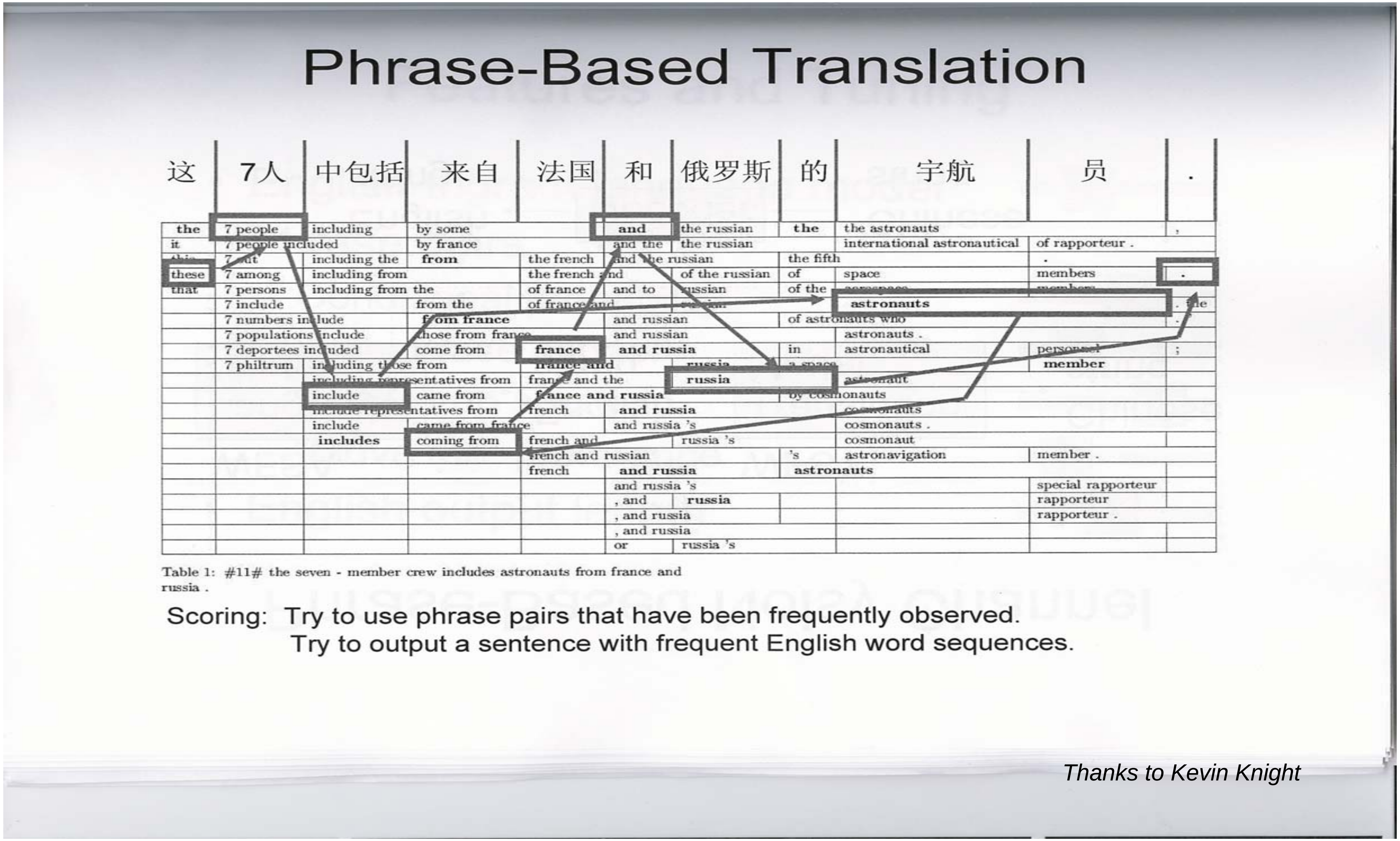
•Add another hypothesis



Hypothesis Expansion

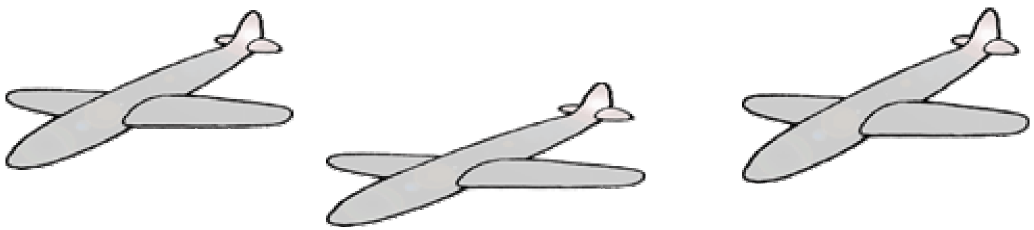
- … until all foreign words covered.
- Find best hypothesis that covers all foreign words
- Backtrack to read off translation
- Problem: Adding more hypotheses causes search space to explode—decoding is NP-complete [Knight 99]
- Solutions:
  - Hypothesis <u>recombination</u>: different paths lead to the same partial translation—risk free!
  - Threshold <u>pruning</u>—risky! (integrated with future cost estimation …)
- Run Pharaoh (or Moses) with the trace on (‘-t’ switch)

Decoding is a Complex Process!



MT Evaluation

- Source only!
- Manual:
  - Subjective Sentence Error Rates
  - Correct/Incorrect
  - Error categorization
- Objective Usage Testing



- Automatic:
  - Exact Match (SER), WER, BLEU, NIST, GTM, Meteor etc.

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 Metric

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!)

NB, 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 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 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 in Theory

- Proposed by IBM's SMT group (Papineni et al, *ACL-2002*)
- Widely used in MT evaluations
  - DARPA TIDES MT evaluation (www.darpa.mil/ipto/programs/tides/strategy.htm)
  - IWSLT evaluation (www.slt.atr.co.jp/IWSLT2004/)
  - TC-Star (www.tc-star.org/)
- BLEU Metric:

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} w_n \log p_n)$$

  - $P_{n:}$ Modified n-gram precision
  - Geometric mean of p_1, p_2,..p_n
  - *BP*: Brevity penalty (*c*=length of MT hypothesis, *r*=length of reference)

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \le r \end{cases}$$

  - Usually, N=4 and $w_n=1/N$.

BLEU in Practice

MT Hypothesis: ***the gunman was shot dead by police .***

- Ref 1: The gunman was shot to death by the police .
- Ref 2: The gunman was shot to death by the police .
- Ref 3: Police killed the gunman .
- Ref 4: The gunman was shot dead by the police .

- Precision: p_1=1.0(8/8) p_2=0.86(6/7) p_3=0.67(4/6) p_4=0.6 (3/5)
- Brevity Penalty: c=8, r=9, BP=0.8825
- Final Score: $\sqrt[4]{1 \times 0.86 \times 0.67 \times 0.6} \times 0.8825 = 0.68$

Sample BLEU Performance

Reference: George Bush will often take a holiday in Crawford Texas

---

1. George Bush will often take a holiday in Crawford Texas (1.000)
2. Bush will often holiday in Texas (0.4611)
3. Bush will often holiday in Crawford Texas (0.6363)
4. George Bush will often holiday in Crawford Texas (0.7490)
5. George Bush will not often vacation in Texas (0.4491)
6. George Bush will not often take a holiday in Crawford Texas (0.9129)

Content of ‘gold standard’ matters!

Which is better?

1. George Bush often takes a holiday in Crawford Texas
2. Holiday often Bush a takes George in Crawford Texas

What would BLEU say (assume max. bigrams important)?

What if human reference was:

The President frequently makes his vacation in Crawford Texas.

Which is better *now*?

Content of 'gold standard' matters! (2)

Sometimes, the reference translation is impossible for *any* MT system (current or future) to match:

From Canadian Hansards:

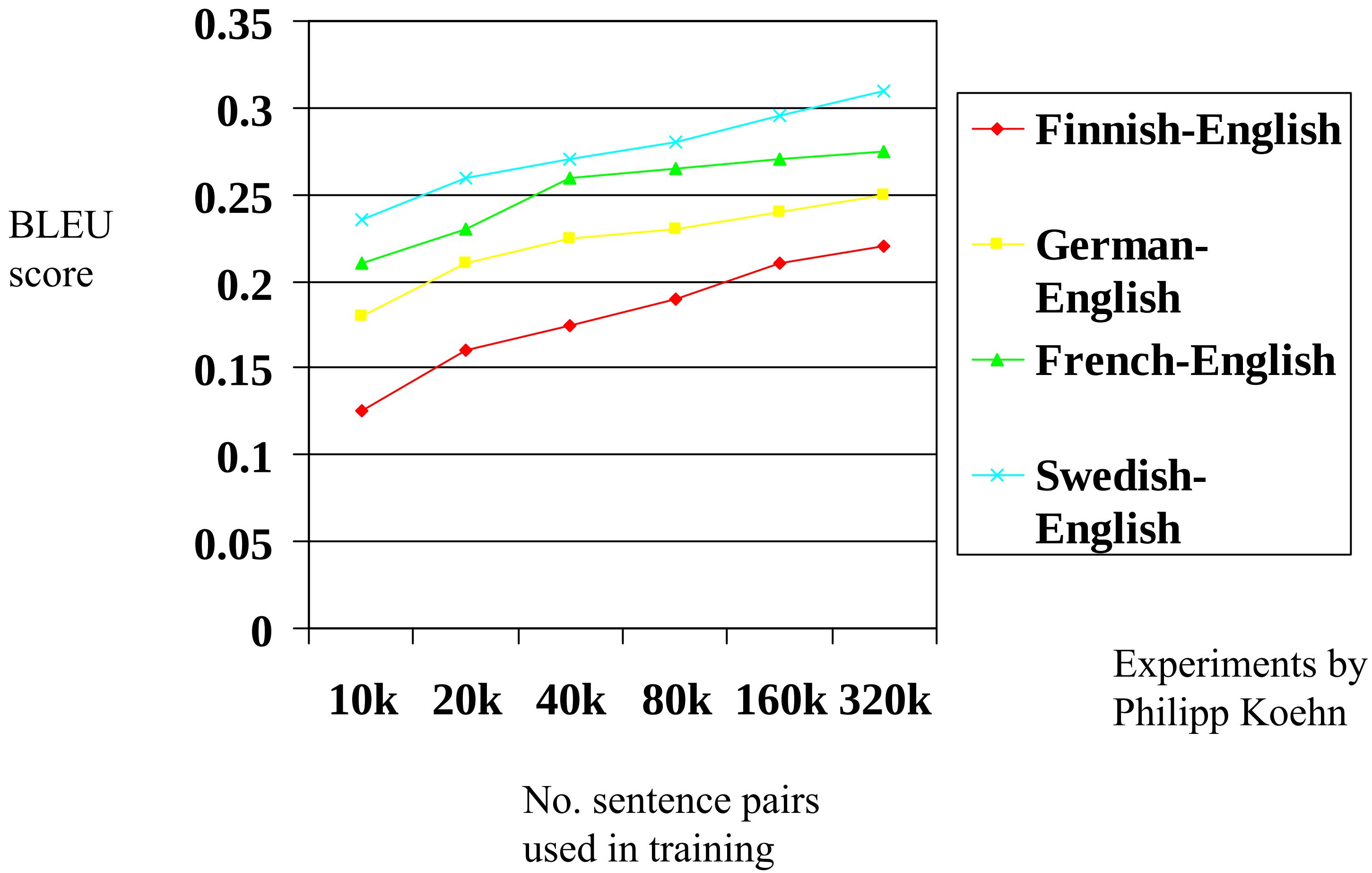
Again, this was voted down by the Liberal majority =>

Malheureusement, encore une fois, la majorité libérale l' a rejeté

[Unfortunately, still one time, the majority liberal it has rejected]

Of course, human translators are quite entitled to do this sort of thing, and do so all the time …

Correlation between BLEU score and Training Set Size?



Experiments by Philipp Koehn

Problems with BLEU

1. It can be easy to look good (cf. output from current ‘state-of-the-art’ SMT systems)
2. Not currently very sensitive to global syntactic structure (disputable)
3. Doesn’t care about nature of untranslated words:
   - *gave it to Bush*
   - *gave it at Bush*
   - *gave it to rhododendron*
4. As MT improves (?!), BLEU won’t be ‘good enough’

Problems with *using* BLEU

- Not designed to test individual sentences
- Not meant to compare different MT systems

<u>Extremely</u> useful tool for system <u>developers!</u>

Q: what/who is evaluation for?

cf. [Callison-Burch et al., *EACL-06*]

Newer Evaluation Metrics

- P&R (GTM: Turian et al., *MT-Summit 03*)
- RED (Akiba et al., *MT-Summit 01*) [based on edit distance, cf. WER/PER ...]
- ORANGE (Lin & Och *COLING-04*)
- Classification by Learning (Kulesza & Shieber *TMI-04*)
- Meteor (Banerjee & Lavie, *ACL-05*)
- TER (Snover et al., *AMTA-06*)

Other Places to Look

- BLEU/NIST: www.nist.gov/speech/tests/mt/resources/scoring.htm
- GTM: nlp.cs.nyu.edu/GTM/
- EAGLES: www.issco.unige.ch/ewg95/ewg95.html
- FEMTI: www.isi.edu/natural-language/mteval/
- MT Summit/LREC workshops etc etc ...

=> MT Evaluation is (one of) the flavour(s) of the month ...

Is MT-Eval for people who can't do MT?

- I used to say so (somewhat mischievously), but some groups that have come up with MT-Eval metrics include:
  - Aachen (Ney)
  - Google (Och)
  - CMU (Lavie, Vogel)
  - NYU (Melamed)
  - Edinburgh (Koehn)
  - Maryland (Dorr)

Is MT-Eval for people who can't do MT?

- I used to say so (somewhat mischievously), but some groups that have come up with MT-Eval metrics include:
  - Aachen (Ney)
  - Google (Och)
  - CMU (Lavie, Vogel)
  - NYU (Melamed)
  - Edinburgh (Koehn)
  - Maryland (Dorr)
  - DCU (Way)

End of Part 1

... But I hope that’s enough to get you started/interested in SMT...

Thanks ... and over to Hany!

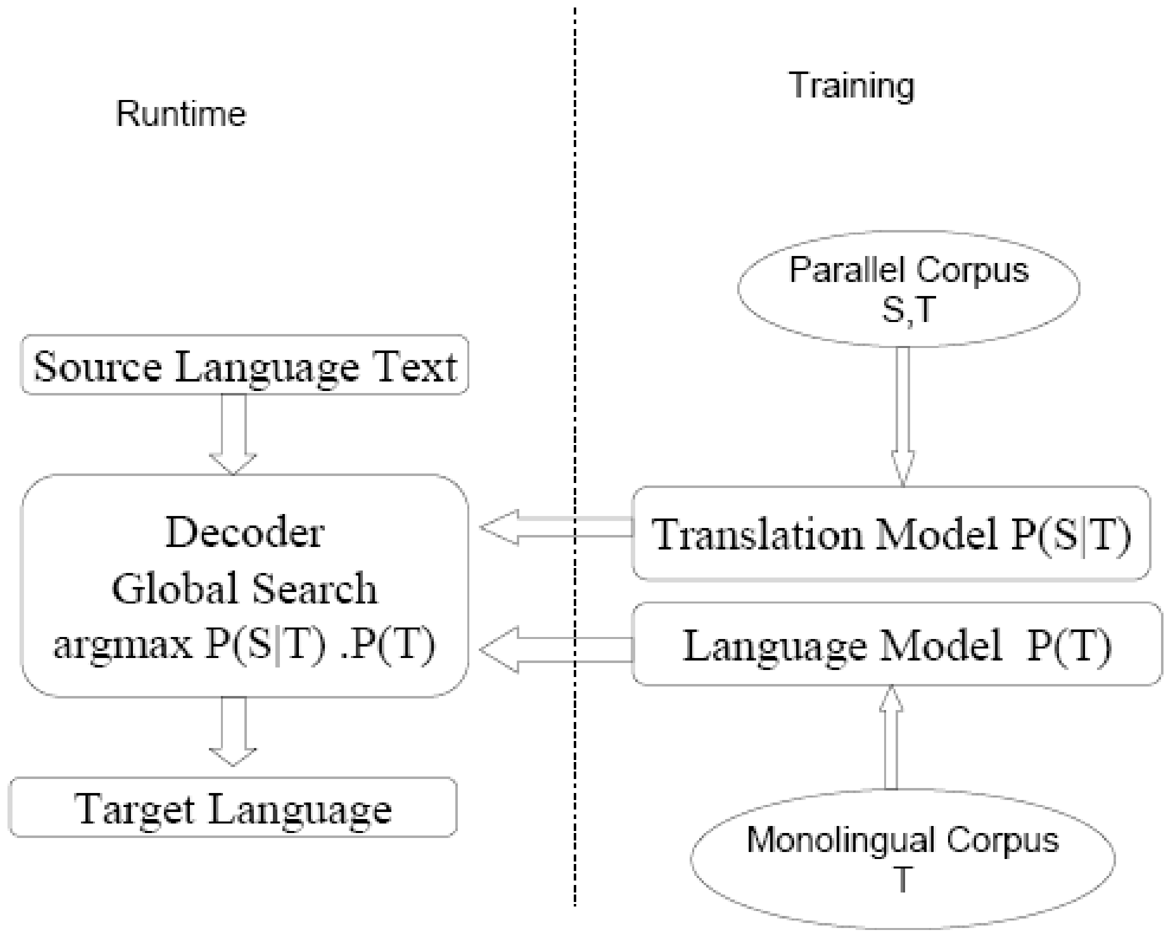
SMT Tutorial – Part2

Andy Way
DCU

Hany Hassan
IBM

Outline

- Phrase-based SMT
  - Log-Linear models & parameters estimation
  - Re-ordering techniques
  - Factored Translation Model
- Advanced Topics:
  - Direct Translation Model
  - Syntax support for SMT
- How to start building your own SMT system?


Runtime
Training
Parallel Corpus
S,T
Source Language Text
Decoder
Global Search
argmax P(S|T) .P(T)
Translation Model P(S|T)
Language Model P(T)
Target Language
Monolingual Corpus
T

Phrase-based SMT
Log-Linear Model

- IBM Models deploys three components:
  - Translation model, Language Model and Distortion model

$$P_{tm} * P_{lm} * P_{dist}$$

- This can be represented as weighted components:

$$P^{\lambda_1}{}_{tm} * P^{\lambda_2}{}_{lm} * P^{\lambda_3}{}_{dist}$$

- Motivated by the need to add new components:

$$\log \prod_i P_i = \sum_i \lambda_i \log P_i$$

Log-Linear model components /features

- Many different knowledge sources useful
  - Phrase translation model
  - Word translation model
  - Reordering (distortion) model
  - Word drop feature
  - Language models
  - Additional linguistics features (i.e. POS)
  - Any feature you can think could be useful

State of-the-art Features

- Source-Target phrase translation
- Target-Source phrase translation
- Source-Target word translation
- Target-Source word translation
- Distortion model
- N-gram Language Model
- Word/phrase deletion penalty

Log-linear models overview

$$\log \prod_i P_i = \sum_i \lambda_i \log P_i \quad \Longrightarrow \quad P = \exp\left(\sum_i \lambda_i \log P_i\right)$$

Log-linear Models

Maximum Entropy Models

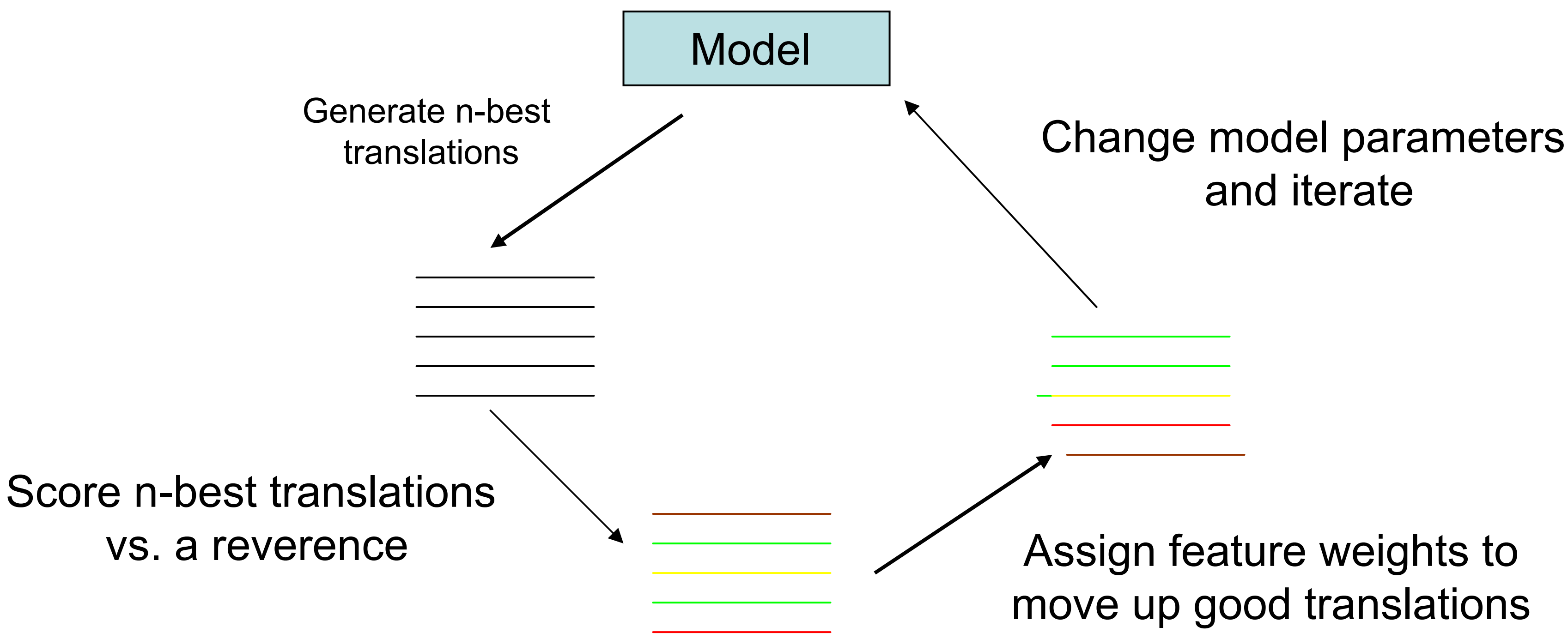
- Log-linear models
  - Heuristic (less optimal) estimation
- Few number of features (< 10)
  - Computationally inexpensive

- Optimal estimation approaches
- Very large number of features (millions)
  - Computationally expensive

Phrase-based SMT was in early development stages
Researchers opted for computationally affordable solution
Still long way to go at that time

Log-linear Model Estimation

- Minimum Error Rate Training (MERT)



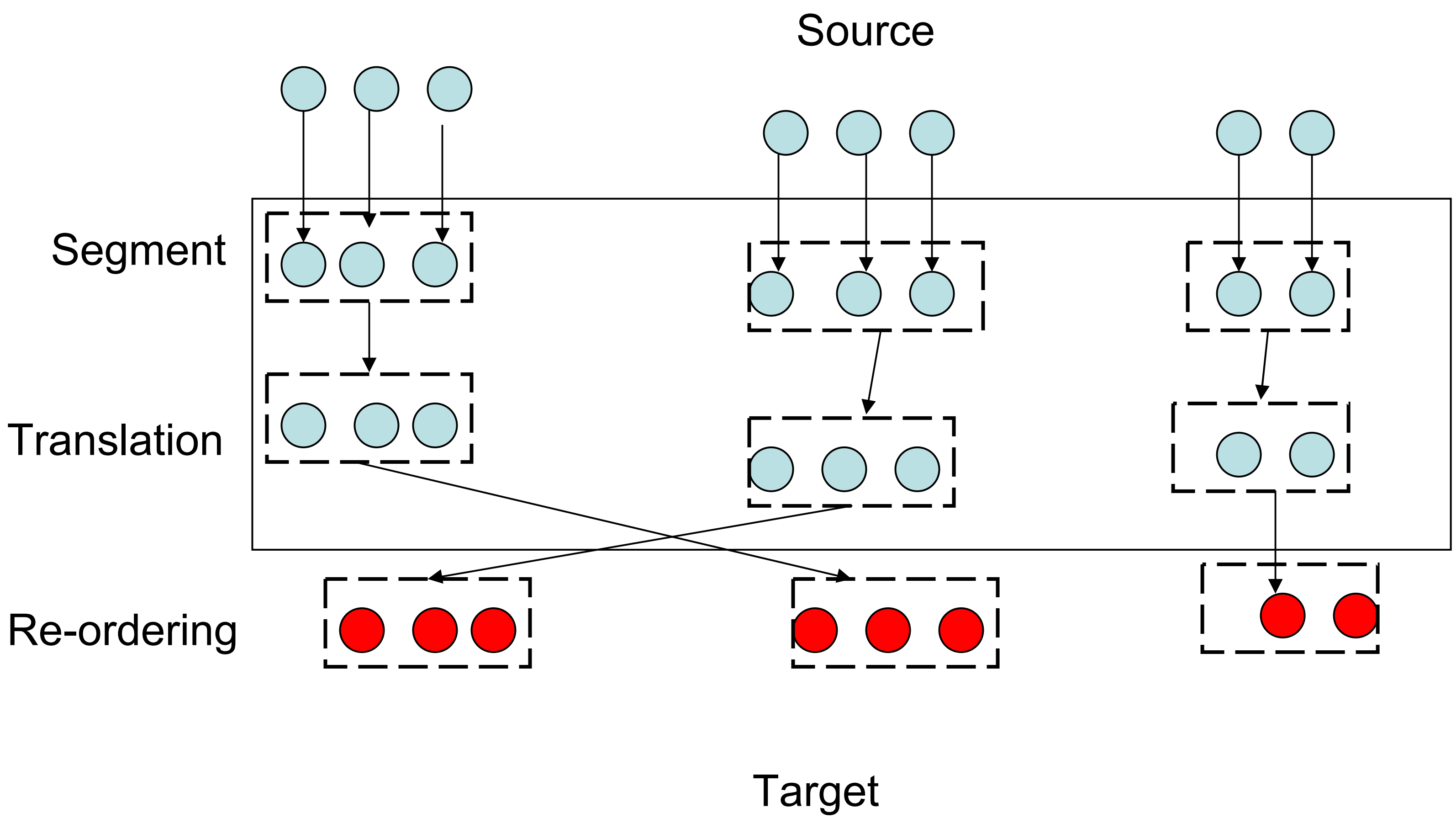
Log-Linear models

- Pros:
  - Proved success and dominated Phrase-based SMT for years
  - Easy to estimate
  - Available open source tools for estimation
- Cons:
  - No optimal estimation
  - Handle few number of features (in the order of ten)
  - Feature weights assigned to the whole feature at once
  - No inter-dependency between features instances

Outline

- Phrase-based SMT
  - Log-Linear models & parameters estimation
  - Re-ordering techniques
  - Factored Translation Model
- Advanced Topics:
  - Discriminative SMT models
  - Syntax support for SMT
- How to start building your own SMT system?

Re-ordering for Phrase-based SMT



Re-ordering

أستقبل الوزير مسئولين اقتصاديين سعوديين

Monotone → met minister officials economic saudi

Re-ordering → Minister met saudi economic officials

Monotone translation with pre-processing

أستقبل الوزير مسئولين اقتصاديين سعوديين

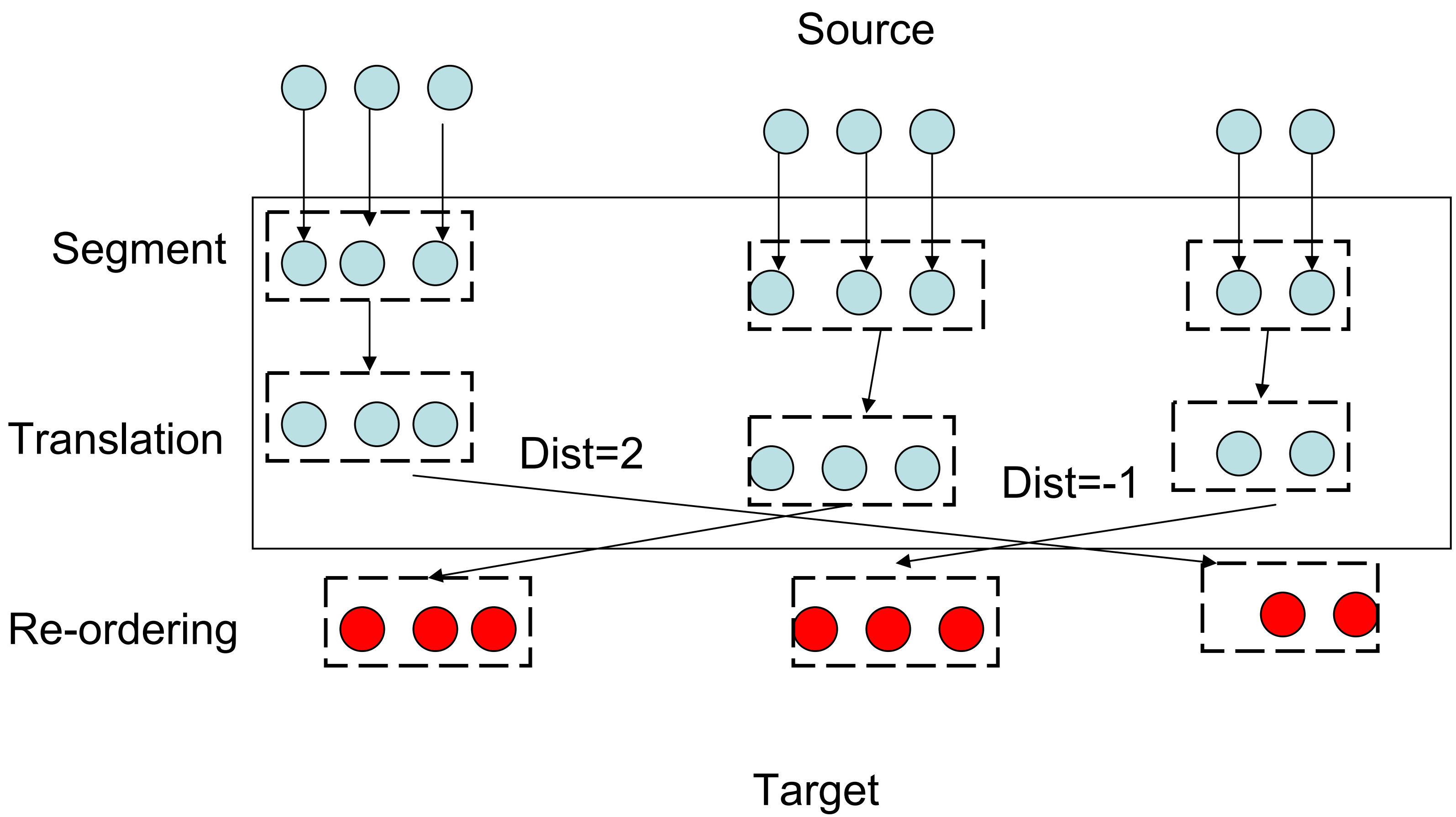
Reorder source to match target → الوزير أستقبل سعوديين اقتصاديين مسئولين

Monotone Decoding → Minister met saudi economic officials

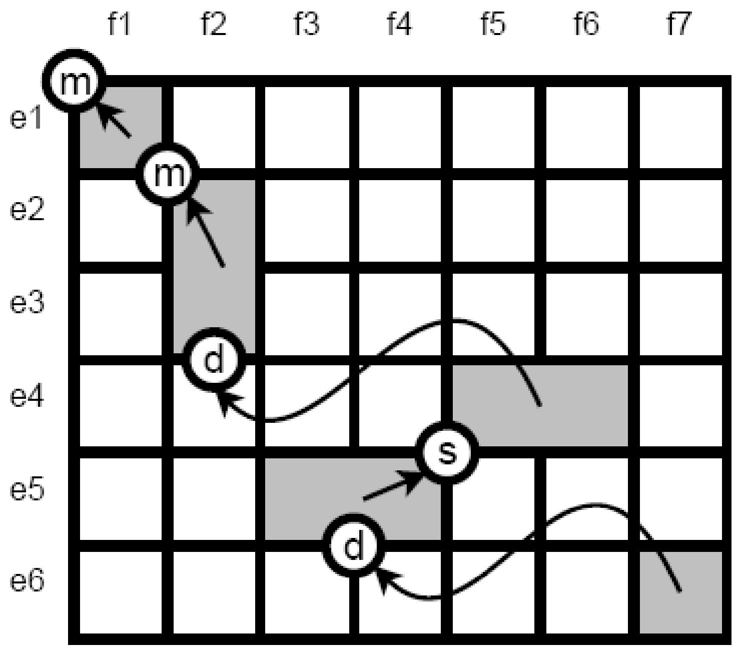
Linear re-ordering

- Model the movement distance
- Independent of the words , phrases and the context
- A weak re-ordering model
- Penalize long movements

Linear Re-ordering for Phrase-based SMT



Lexicalized re-ordering



Three orientation types: monotone, swap, discontinuous

- Probability p(swap|e, f) depends on foreign (and English) phrase involved

More re-ordering techniques

- POS based re-ordering
- Syntax based re-ordering
- etc.

Lexical Reordering is doing a good job
n-gram language models limits the reordering capabilities
Seeking better language modeling techniques to pick the best re-ordering
Syntax-based language models ?

Outline:

- Phrase-based SMT
  - Log-Linear models & parameters estimation
  - Re-ordering techniques
  - Factored Translation Model
- Advanced Topics:
  - Discriminative SMT models
  - Syntax support for SMT
- How to start building your own SMT system?

Factored Translation Model

- Factored Translation Models
  - Factored representation of words
    - surface
    - stem
    - part-of-speech
    - morphology
    - word class
- Generalization, e.g. by translating stems, not surface forms
- Additional information within model (using syntax for reordering, language modeling)

Factored Translation Model

Decomposing Translation: Example

- **Translating** stem and syntactic information **separately**

$$\begin{bmatrix} \text{stem} \end{bmatrix} \Rightarrow \begin{bmatrix} \text{stem} \end{bmatrix}$$

$$\begin{bmatrix} \text{part-of-speech} \\ \text{morphology} \end{bmatrix} \Rightarrow \begin{bmatrix} \text{part-of-speech} \\ \text{morphology} \end{bmatrix}$$

- **Generate surface** form on target side

$$\begin{bmatrix} \text{surface} \end{bmatrix}$$
$$\Uparrow$$
$$\begin{bmatrix} \text{stem} \\ \text{part-of-speech} \\ \text{morphology} \end{bmatrix}$$

Factored Translation

- Pros:
  - Provides a framework to deploy various knowledge sources
  - Implemented in Moses framework
- Cons:
  - Few number of features (<10)
  - No adequate estimation and modeling
  - Not correlating various features
  - Redundant and overlapping features

Outline

- Phrase-based SMT
  - Log-Linear models & parameters estimation
  - Re-ordering techniques
  - Factored Translation Model
- Advanced Topics:
  - Direct Translation Model
  - Syntax support for SMT
- How to start building your own SMT system?

Direct Translation Model

- Why?
  - Provides a framework to deploy various knowledge sources
  - Easy to understand classification approach
  - Very large number of features
  - estimation and modeling
  - Automatically correlating various features
  - Minimal no redundant phrase table

A Classification Viewpoint

- Machine Translation can be viewed as a sequence of tagging decisions
- Classifier
  - MaxEnt
  - ...
- Required:
  - History (Flip of a coin, classifiable action)
  - Futures (An outcome)
- Nice to have:
  - Relevant Features

Log-linear models overview

$$\log \prod_i P_i = \sum_i \lambda_i \log P_i \quad \Rightarrow \quad P = \exp\left(\sum_i \lambda_i \log P_i\right)$$

Log-linear Models

Maximum Entropy Models

- Log-linear models
    - Heuristic (less optimal) estimation
- Few number of features (< 10)
    - Computationally inexpensive

- Maxent models
- Optimal estimation approaches
- Very large number of features (millions)
    - Computationally expensive

Phrase-based SMT is more mature now
Researchers started to hit the upper limits of the log linear models capabilities
Computational power increases remarkably

DTM

- The model:

$$p(t_i, j \mid t_{i-2}^{i-1}, s_{a_i-1}^{a_i+1}) = \frac{1}{Z} P_0 e^{\sum_i \lambda_i \phi(t_i, j, t_{i-}, s_{a_i-1}^{a_i+1})}$$

DTM: Generation Story

- Given a source sequence,
  1. Choose a source position
  2. Choose a translation string
  3. Mark source position as covered
  4. Iterate from step 1, till all positions are covered

Not much different from a phrase based decoder…

DTM Features

- Features Types
  - Lexical
  - Segmentation
  - Lexical Context
  - Part of speech
  - Coverage
  - ...

Minimal Phrase Table with Hierarchical Structures

Flat phrases
Large Redundancy
Large Space

Saudi economic officials (11)

Saudi economic officials **(11)**

Saudi political officials (8)

Saudi economic officials **(7)**

Hierarchical Phrases
Minimal Phrase table
Minimal redundancy

official in ↔ مسئولا X في
Saudi X official ↔ مسئولا X سعوديا
meets Saudi ↔ يستقبل X سعوديا

Outline

- Phrase-based SMT
  - Log-Linear models & parameters estimation
  - Re-ordering techniques
  - Factored Translation Model
- Advanced Topics:
  - Discriminative SMT models
  - Syntax support for SMT
- How to start building your own SMT system?

Why syntax

- Syntax can help Phrase-based SMT in:
  - Producing more fluent translation
  - Syntax –aware re-ordering

Source: بوغتا ٤-١٢ (ا ف ب) - ذكر مراسل وكالة فرانس برس ان زعيم كارتل كالي (جنوب غرب) جيلبرتو رودريغس اوريهول ، احد اكبر مهربي المخدرات في العالم ، سلم مساء الجمعة الي الولايات المتحدة .

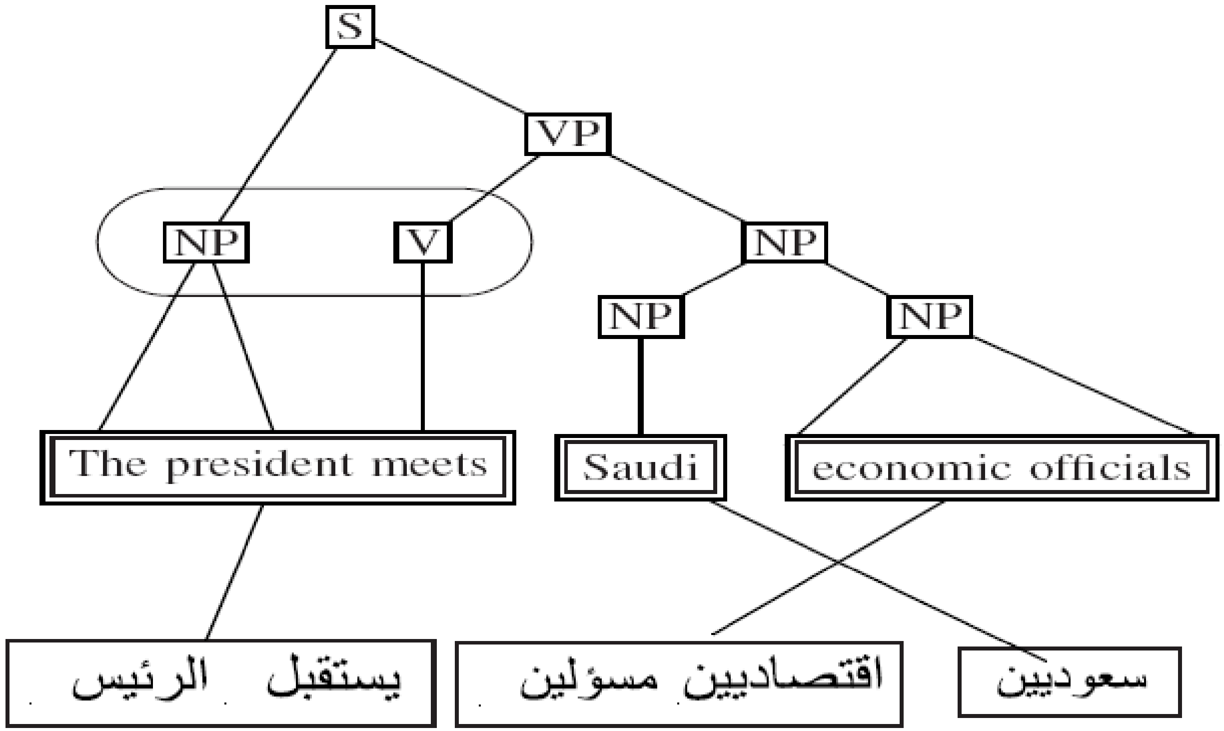
Reference: *Bogota 12-4 (AFP) - An Agence France-Presse correspondent reported th at Cali cartel boss (south-west) Gilberto Rodriguez Orejuela, one of the biggest drug traffickers in the world, was handed over to the United States on Friday e vening.*

Baseline: *Bogota 4-12 (afp) - according to an Agence France Presse correspondent that cali cartel leader (southwest) , gilberto rodriguez orejuela , one of the biggest drug traffickers in the world , surrendered friday night to the united states .*

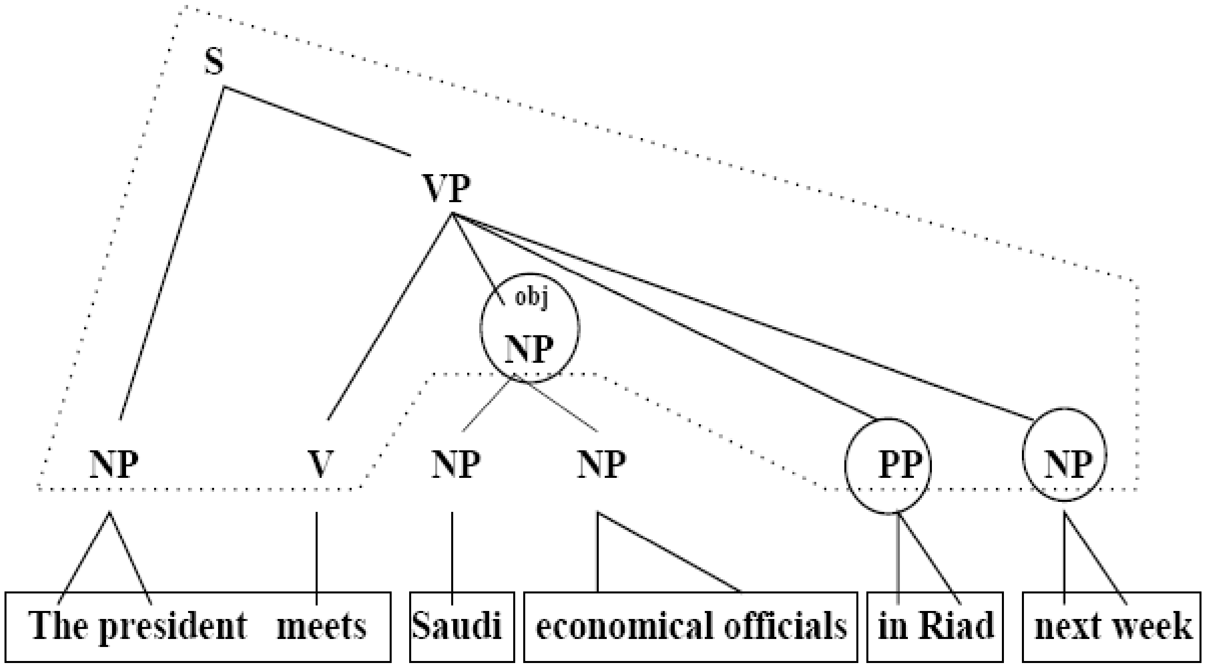
Can linguistic syntax improve PBSMT?

- Early work tried to impose syntactic constituents on phrase extraction with no success
- Hierarchical Phrase structure
  - Allows for hierarchical phrases
  - Handles a range of reordering problems
  - The syntax induced is not linguistically motivated.
- Syntactified target phrases
  - Induces millions of xRs rules from parallel corpus
  - Mismatch between constituent (xRs) and phrase
  - Subtrees for phrases: leads to spurious ambiguity in phrase table
- Do subtrees/constituents fit well with phrases?

Subtrees mismatch phrases

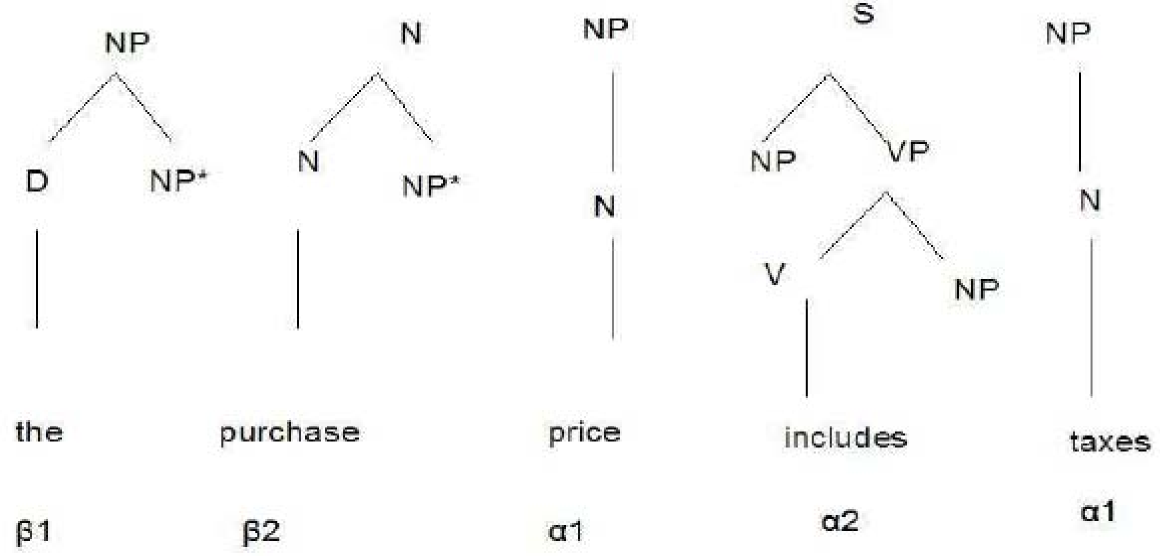


Redundancy

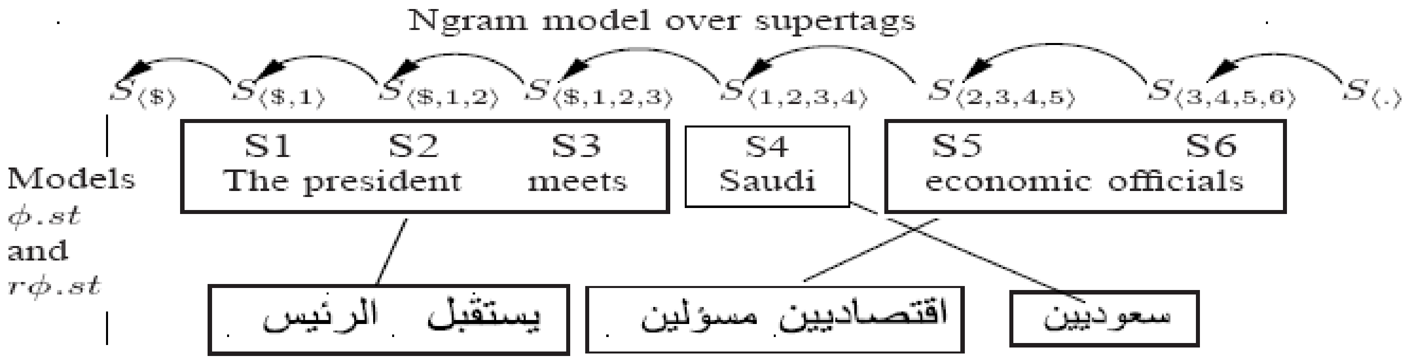


S
VP
obj
NP
NP
V
NP
NP
PP
NP
The president meets
Saudi
economical officials
in Riad
next week

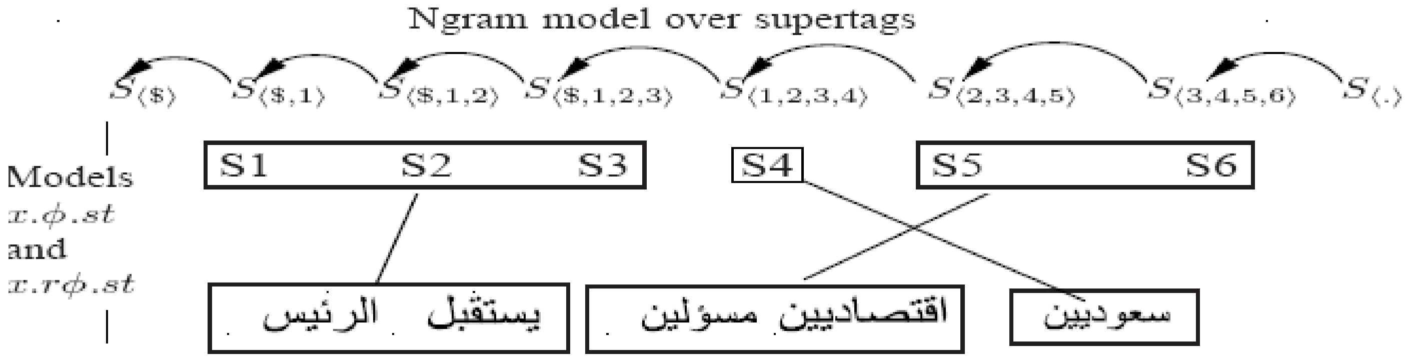
Lexical Syntax



Supertagged Phrase-based SMT



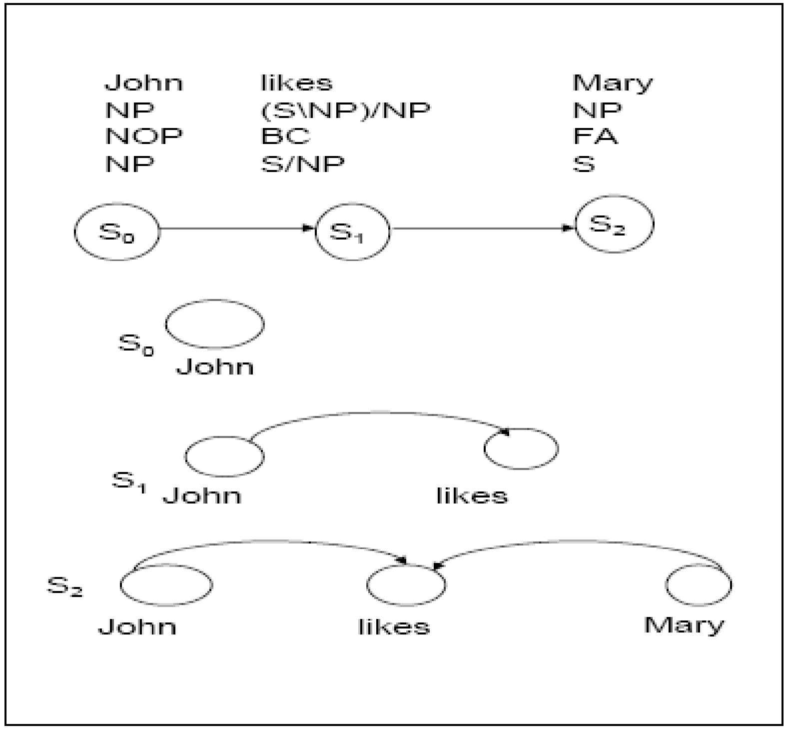
Unsmoothed Supertagged System



“almost” parsing for SMT

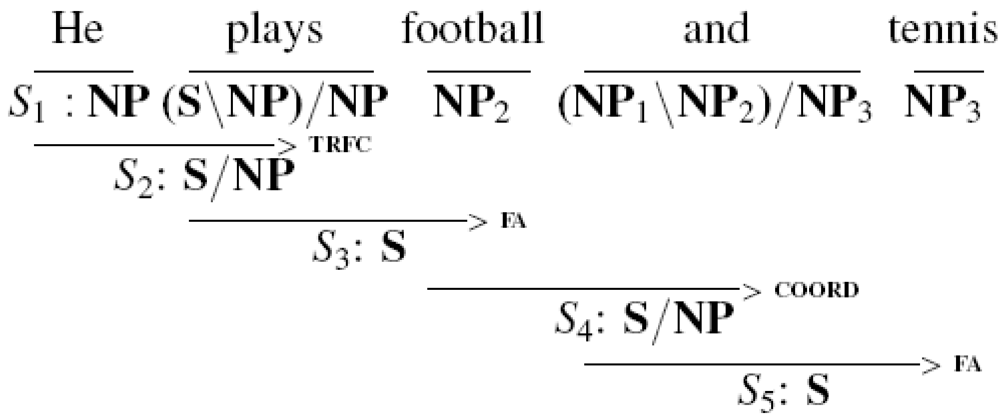
- Phrases with supertags information
- Translation models to handle both lexical and supertagged phrases
- Lexical language model
- Supertagged /Syntactic language model
- Very efficient linear decoding
- Very good improvement

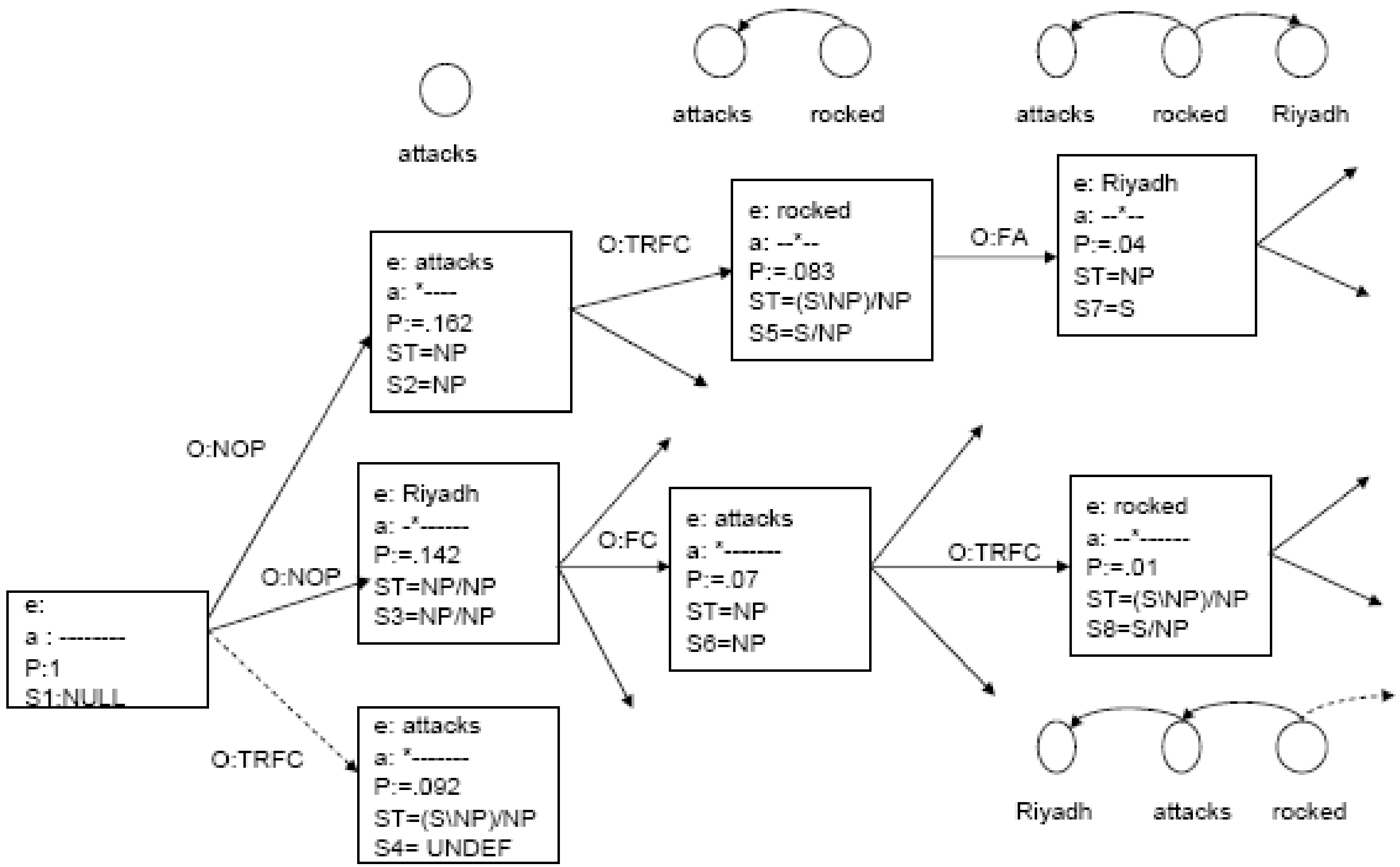
Incremental dependency parsing using lexical syntax

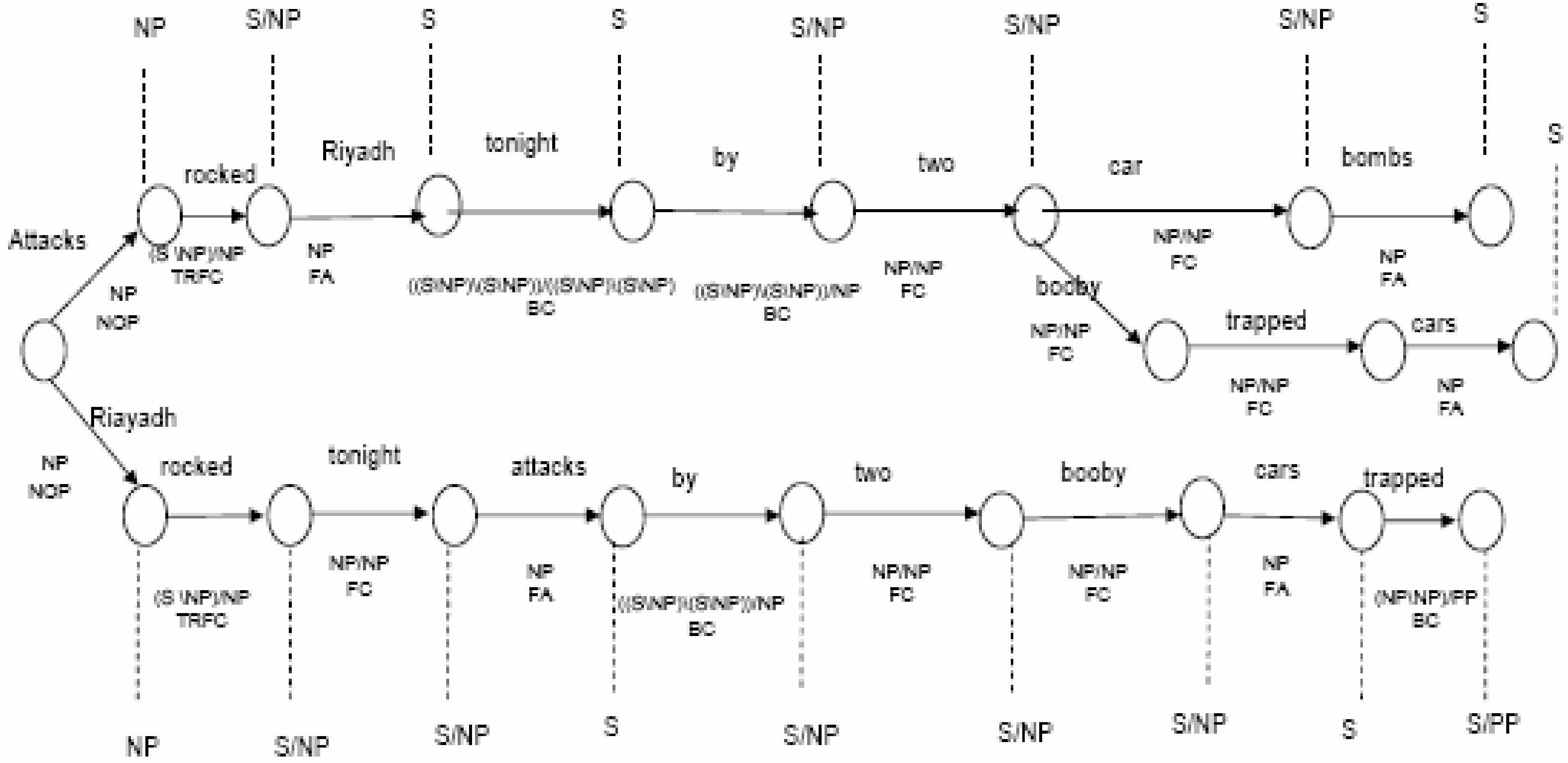


$$P(W,S) = \prod_{i=1}^{n} \overbrace{P(w_i|W_{i-1}S_{i-1})}^{Word\ Predictor} . \overbrace{P(st_i|W_i)}^{Supertagger} . \overbrace{P(o_i|W_i, S_{i-1}, ST_i)}^{Operator\ Tagger}$$

Incremental CCG




attacks
attacks
rocked
attacks
rocked
Riyadh
e:
a : --------
P:1
S1:NULL
O:NOP
O:NOP
O:TRFC
e: attacks
a: *----
P:=.162
ST=NP
S2=NP
e: Riyadh
a: -*------
P:=.142
ST=NP/NP
S3=NP/NP
e: attacks
a: *-------
P:=.092
ST=(S\NP)/NP
S4= UNDEF
O:TRFC
e: rocked
a: --*--
P:=.083
ST=(S\NP)/NP
S5=S/NP
O:FA
e: Riyadh
a: --*--
P:=.04
ST=NP
S7=S
O:FC
e: attacks
a: *-------
P:=.07
ST=NP
S6=NP
O:TRFC
e: rocked
a: --*------
P:=.01
ST=(S\NP)/NP
S8=S/NP
Riyadh
attacks
rocked


NP
S/NP
S
S
S/NP
S/NP
S/NP
S
S
Attacks
NP
NOP
rocked
(S \NP)/NP
TRFC
Riyadh
NP
FA
tonight
((S\NP)\(S\NP))/((S\NP)\(S\NP))
BC
by
((S\NP)\(S\NP))/NP
BC
two
NP/NP
FC
car
NP/NP
FC
bombs
NP
FA
booby
NP/NP
FC
trapped
NP/NP
FC
cars
NP
FA
Riayadh
NP
NOP
rocked
(S \NP)/NP
TRFC
tonight
NP/NP
FC
attacks
NP
FA
by
((S\NP)\(S\NP))/NP
BC
two
NP/NP
FC
booby
NP/NP
FC
cars
NP
FA
trapped
(NP\NP)/PP
BC
NP
S/NP
S/NP
S
S/NP
S/NP
S/NP
S
S/PP

Syntax effect

Source: وخضع بعد ذلك لفحوصات اجراهَا احد اطباء الشرطة

Reference: *He then underwent medical examinations by a police doctor*
Baseline: *He was subjected after that tests conducted by doctors of the p*
DDTM: *Then he underwent tests conducted by doctors of the police .*

Source: وقد هز الرياض مساء اليوم هجومان بسيارتين مفخختين

Reference: *Riyadh was rocked tonight by two car bomb attacks..*
Baseline: *Riyadh rocked today night attacks by two booby - trapped car*
DDTM: *Attacks rocked Riyadh today evening in two car bombs.*

Where to go from here?

- Open source frameworks
  - Word based aligner : Giza++
  - Open source phrase-based system training and decoding: Moses
  - Language Model tools : SRILM
  - Syntax-based SMT system: SAMT
- Parallel Data
  - LDC data (Arabic, English, Chinese, etc)
  - Europal data (European Languages)
- Monolingual data
  - LDC data
  - Google web n-gram data
- Pre-processing tools
  - OpenNLP, CADIM, AMIRA, ..
- Parsers
  - Bikel's parser, CCG parser, etc