The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign

Yves Lepage Adrien Lardilleux

GREYC, University of Caen, France

Monday, October 15th

Outline

The system

- Evolution of the ALEPH machine translation system that participated in the IWSLT 2005 campaign [Lepage & Denoual, 2005].
- IWSLT 2007: participation to the two classical tasks (Japanese to English, Arabic to English).
- ALEPH is a pure example-based system that exploits proportional analogies between character strings (analogies of form):

$$A : B :: C : D$$

Examples of analogies

to eat : *he eats* :: *to drink* : *he drinks*

cup of coffee : *cup of tea* :: *have a coffee* : *have a tea*

one dollar : *two dollars* :: *one piece* : *two pieces*

Principle of corresponding analogies

↓

月曜日に会い
ましょう。
/getuyoubinikaiimasyou./

Principle of corresponding analogies

↓

月曜日に会いましょう。 : x :: 月曜日に御会いしましょう。 : 明日御会いしましょう。

/getuyoubinikaiimasyou./ /getuyoubinioaisimasyou./ /asitaoaisimasyou./

Principle of corresponding analogies

↓

月曜日に会いましょう。 : x :: 月曜日に御会いしましょう。 : 明日御会いしましょう。

/getuyoubinikaiimasyou./ /getuyoubinioaisimasyou./ /asitaoaisimasyou./

⇓

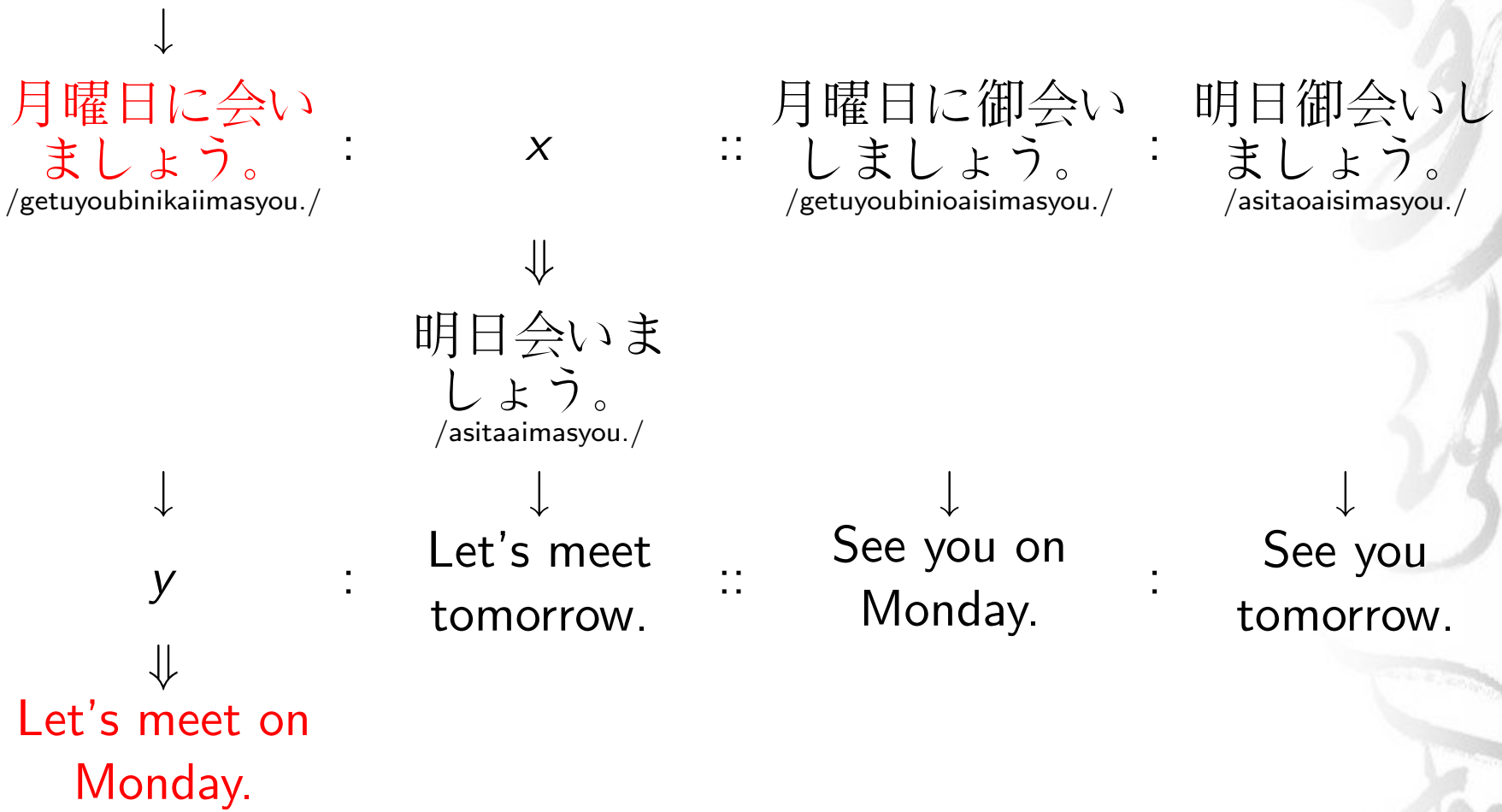
明日会いましょう。

/asitaaimasyou./

Principle of corresponding analogies

↓						
月曜日に会いましょう。 /getuyoubinikaiimasyou./	:	x	::	月曜日に御会いしましょう。 /getuyoubinioaisimasyou./	:	明日御会いしましょう。 /asitaoaisimasyou./
		⇓				
		明日会いましょう。 /asitaaimasyou./				
↓		↓		↓		↓
y	:	Let's meet tomorrow.	::	See you on Monday.	:	See you tomorrow.

Principle of corresponding analogies



Main issue: size of the training data

20,000 training sentences are not sufficient to get any translations (analogies are not numerous enough).

When unable to translate by analogy, the engine backs off to the basic behavior of a translation memory.

IWSLT 2005: 140,000 extra sentences from the BTEC were used.

IWSLT 2007: cope with the 20,000 or 40,000 provided sentences only!

→ Enhancement needed

2 ENHANCEMENTS PROPOSED

1. Inflate the training data by adding sub-sentential alignments.
2. Use of a heuristic to increase the number of successfully solved analogical equations.

Need for sub-sentential alignments

The number of analogies between chunks tends to be the square of the number of analogies between sentences [Lepage & al., last week].

→ Expand the data with "close-to-chunk" sub-sentential alignments:

- words
- chunks:

  Japanese: chunking is based on markers (→ *bunsetsus*)

。	、	の	で	へ	に	を	は	が	から	ました
/./	/,/	/no/	/de/	/e/	/ni/	/wo/	/wa/	/ga/	/kara/	/-masita/

  Arabic: a form separated by two spaces corresponds to some extent to the notion of a chunk in English

Alignment method

Previous research revealed that the use of *hapax legomena* (frequency = 1) could yield good alignment results [Lardilleux & Lepage, last week].

→ Experimentation of a new alignment method: create subcorpora where the strings to be aligned are artificially made hapaxes.

IF POSSIBLE: align the source and target hapaxes

IF NOT: the strings are not aligned

(ongoing work)

Some alignment results

للأمام، وهو
/lʾamām, whw/
'Go straight, and it'
↔ Go straight,

平土間
/tairadoma/
'stalls'
↔ the stalls

يفتح المصرف
/yftḥ ālmṣrf/
'the bank open'
↔ does the bank open

一時間ほどで
/itizikanhodode/
'about an hour'
↔ about an hour

سياتل
/syātl/
'Seattle'
↔ Seattle

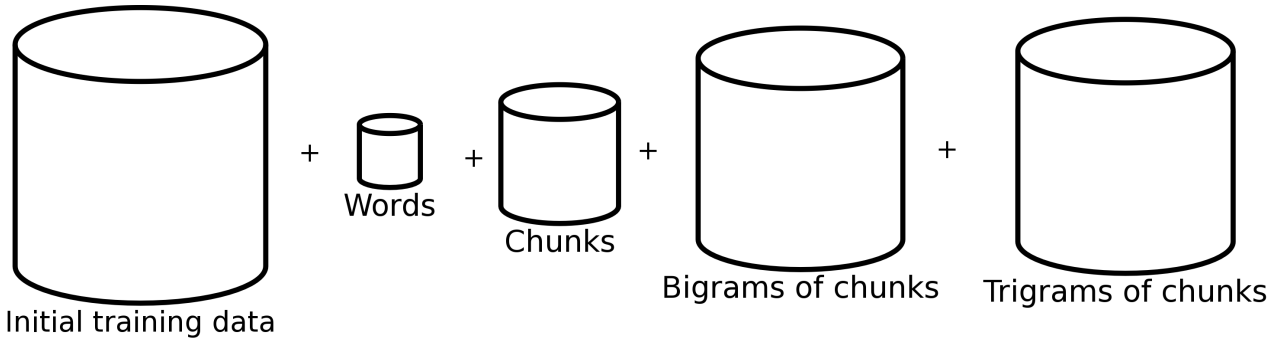
このバスは動物園迄
/konobasuhadoubutuenmade/
'this bus to the zoo'
↔ the zoo?

؟
/?/
'?'
↔ ?

今空いていますか
/imaaiteimasuka/
'is it vacant now'
↔ the toilet vacant now

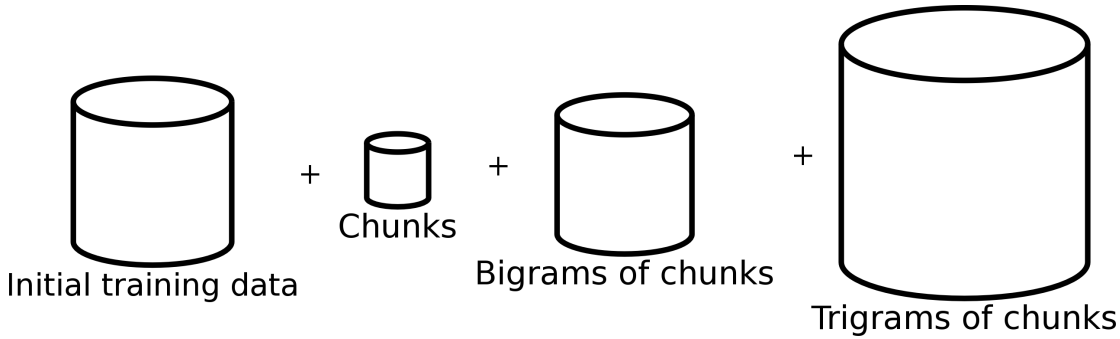
Final data used for translation

Japanese-English:



$\simeq 128,000$ alignments

Arabic-English:



$\simeq 60,000$ alignments

Enhancement by engine improvement

In addition to the previous analogical equation in the source language $(A : x :: C : D)$, we also try:

$$A : B :: C : x$$

where B is close to A, and C is well included in A.

(not in D! Mistake in the paper p.4)

$\rightarrow$ This heuristic has proved to be productive thanks to the expansion of the training data with sub-sentential alignments.

Japanese to English task results

55% of test sentences were found in our training data:

まっすぐ行って下さい。
/massuguittekudasai./
'Go straight on.' ↔ Go straight on.

17% were translated by analogy:

明後日の朝迄に。
/myougonitinoasamadeni./
'By morning, the day after tomorrow.' ↔ The day after tomorrow morning.

28% were not translated (translation memory):

これバラバラに壊れています。
/korebarabaranikowareteimasu./
'This fell apart.' ↔ This is broken.

BLEU score = 0.396 (ranking: 8/9)

Arabic to English task results

7% of test sentences were found in our training data:

أين مقعدي؟
/ʾayn mqʿdy?/
'Where's my seat?'
↔ Where's my seat?

15% were translated by analogy:

كيف تُفضّلها؟
/kyf tufḍlhā?/
'How would you like it?'
↔ How do you like it?

78% were not translated (translation memory):

أشعر بالنعاس.
/ʾašʿr bālnʿās./
'I feel sleepy.'
↔ I feel ill.

BLEU score = 0.329 (ranking: 10/11)

Conclusion

- Major characteristic of this EBMT system: totally endogenous
- Main goal for this year was partially completed:
  more test sentences were translated by analogy
- The two improvements proposed still can be improved!