MULTIUN

A MULTILINGUAL CORPUS FROM UNITED NATION DOCUMENTS

Andreas Eisele, Yu Chen

DFKI GmbH, Saarbrücken, Germany

May 21, 2010

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

WHY MULTIUN?

- Importance of having parallel corpora
  - Building SMT systems
  - Developing cross-lingual applications
  - Propagating linguistic knowledge across languages
- Importance of having corpora with potential to grow
  - Supplying up-to-date lexicons
  - Regular supply of unseen test data
- Importance of having multilingual corpora
  - Providing compact information
  - Allowing triangulation approaches

Existing Corpora

- Multilingual corpora (mostly in European languages)
  - Europarl (11 languages)
  - UMC (English-Czech-Russian)
  - UN Parallel text (English-French-Spanish)
  - JRC-Acquis (23 languages)
  - UN General Assembly Resolutions (6 languages)
- Growing corpora
  - Europarl (Release v5 in Jan. 2010)

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

Overview of MultiUN Corpus

- Languages:
  - Official languages

    English, French, Spanish, Arabic, Russian, Chinese
  - Additional: German
- Size:
  - 463,406 documents encoded in XML
  - 1% in German
- Updates:
  - 01/2000 - 10/2009
  - 10,000 new files added in 6 months

    (between pre-release and release v1)

Official Document System (ODS) of UN

- Files in ODS
  - a file ID, which is unique for each individual file,
  - the language labeled for each document in the ODS,
  - the publication date of the document, and
  - a document *symbol*
    - a unique identifier consisting of numbers and letters for a United Nations document
      e.g. “A/CONF.157/PC/63/Add.4”: document No. 63 from the Preparatory Committee of the World Conference on Human Rights in General Assembly

Collection Process

- Crawling
- Preprocessing
  - Text extraction
- Selection and cleaning
  - Noise filtering
  - Language identification (*TextCat*)
- Formatting
  - Sentence segmentation (*NLTK*)
  - Renaming all files after document symbols and languages
    "*N0831582*" → "*S_AGENDA_5874-ru*"
- Sentence alignment (*hunalign*)
- Common test set
  - 6 months of documents reserved

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

Monolingual Data

Language	Documents	Sentences	Words*
English	96240	17098695	385894793
French	85651	14805529	377242310
Spanish	70509	13052875	352460926
Arabic	65156	11050313	237412090
Russian	77061	13852535	278606813
Chinese	65022	10839473	756108566
German	3763	232225	5848668

∗Only number of characters are counted for Chinese texts.

NUMBER OF DOCUMENT PAIRS

	fr	es	ar	ru	zh	de
en	96240	68314	63257	74053	62815	3643
fr		68014	63193	73973	62738	3632
es			63241	64230	62707	3632
ar				63194	63031	3677
ru					62842	3635
zh						3886

Number of sentence pairs

	fr	es	ar	ru	zh	de
en	12M	11M	8M	6M	9M	156K
fr		11M	9M	8M	9M	153K
es			9M	7M	9M	150K
ar				9M	7M	144K
ru					5M	120K
zh						153K

Number of words (L1:L2)

	fr	es	ar	ru	zh	de
en	302M	267M	215M	164M	220M	5M
	338M	311M	181M	97M	629M	5M
fr		316M	255M	236M	243M	5M
		329M	195M	154M	628M	4M
es			262M	210M	253M	5M
			193M	128M	622M	4M
ar				201M	162M	4M
				197M	574M	4M
ru					68M	3M
					392M	3M
zh						13M
						4M

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

BLEU SCORES OF THE SMT SYSTEMS

Language pairs	Development set	Test set
Spanish-Chinese	33.25	31.35
Chinese-Spanish	40.65	39.08
French-Chinese	29.40	29.94
Chinese-French	34.85	34.66

TRANSLATION EXAMPLE

SOURCE: bsmi 国民议会第二次常会于5月4日至22日举行。

REFERENCE: La deuxième session spéciale de l'Assemblée nationale a eu lieu du 4 au 22 mai.

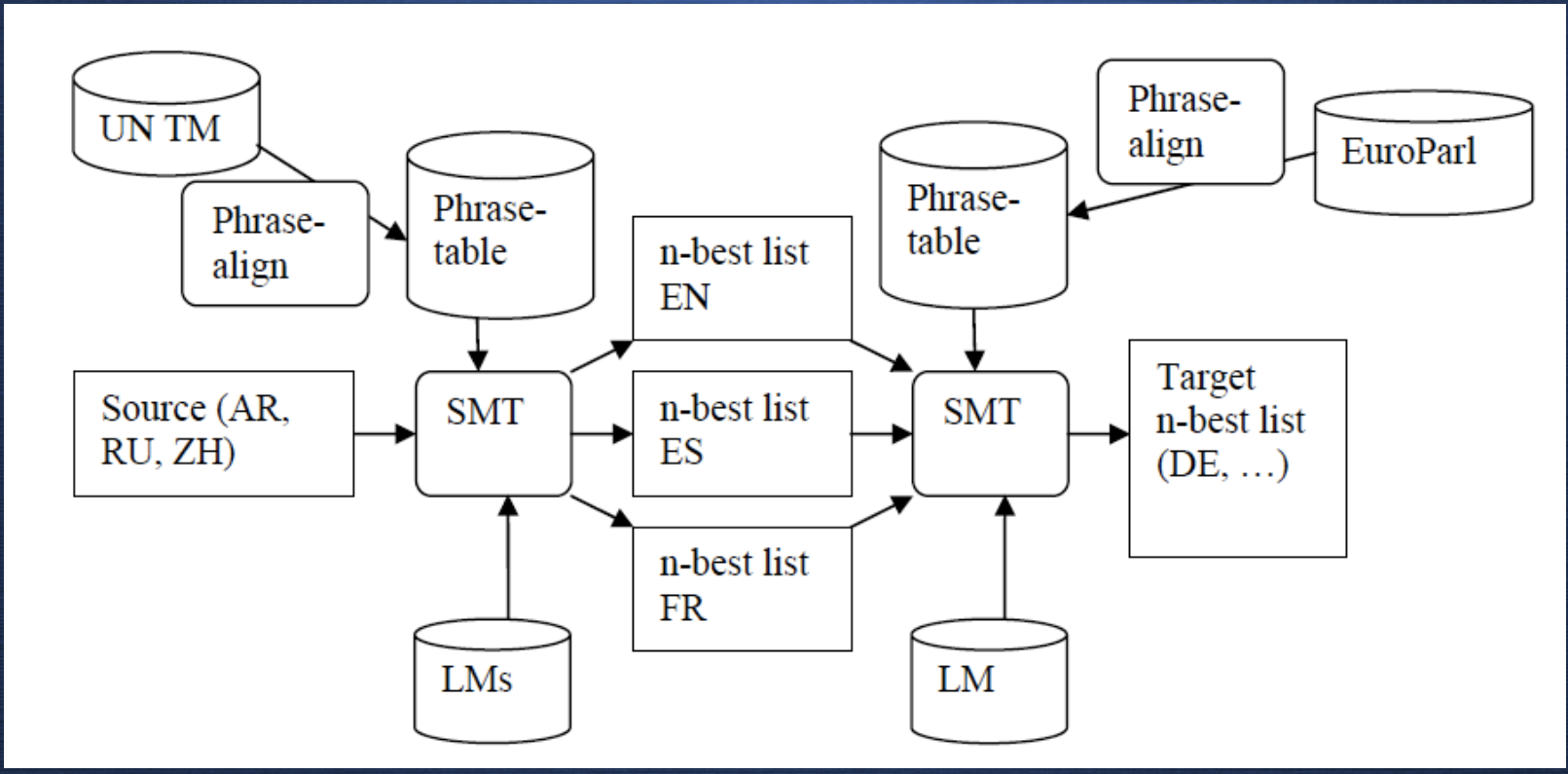
IN-HOUSE: L'Assemblèe nationale à la deuxième réunion ordinaire du 4 au 22 mai .

ONLINE: La deuxième session ordinaire de l'Assemblée nationale le 4 Mai 22 a été tenue.

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

Example of Potential Use



See Eisele 2006 for more details on three-way pivot MT

Outline

- Introduction
- Corpus collection
- Corpus property
- Experiments
- Example of Potential Use
- Conclusion

SUMMARY

- Multilingual parallel corpus extracted from official documents of the United Nations
  - 6+1 languages
  - 463K documents, 81M sentences in total
  - avg. 79K documents, 326M tokens per languages *(excl. Chinese, German)*
- SMT systems constructed with the corpus
  - French↔Chinese
  - Spanish↔Chinese

Future Plan

- Refinements on collection process
- Yearly updates
- Multiway alignments