PANACEA WP5: Parallel Corpus and Derivatives

Technical Meeting
15th – 16th Apr. 2010
ILSP, Athens

Pavel Pecina and Antonio Toral, DCU
Gregor Thurmair, Linguatec

Overview

1. WP5 Tasks & Deliverables
2. Overview of parallel technology tools
3. Parallel copora requirements
4. Survey of language resources
5. Work plan for t4-t14
6. Questions

WP5: Tasks

Input: parallel corpora produced in WP4

Output: language resources for MT in WP7/WP8

Tasks:

WP5.1 Sub-sentential alignment (DCU, ELDA, ILSP)

WP5.2 Bilingual dictionary extraction (DCU, ILSP)

WP5.3 Transfer grammar induction (LT)

WP5: Deliverables

- **D5.1** (*t06*): Report describing the inventory of parallel technology tools to be developed and integrated in PANACEA and the characteristics of the resources to be produced.
- **D5.2** (*t14*) Aligners integrated into the platform, and documentation (scientific paper).
- **D5.3** (*t22*) Parallel, sententially aligned texts, cleaned and prepared for training/building translational models (20—50 million words) combining EN, DE, ES, IT, FR & EL.
- **D5.4** (*t30*) Final version of the Bilingual Dictionary Extractor integrated and documentation.
- **D5.5** (*t30*) Sample of bilingual dictionaries produced: EN—FR and EN—EL for 100K lemmas.
- **D5.6** (*t30*) Final version of the integrated Transfer Rules module, and documentation.
- **D5.7** (*t30*) Sample of transfer rules produced for EN—DE.

Parallel technology tools

- Sub-sentential alignment (WP5.1)
- Billingual dictionary extraction (WP5.2)
- Transfer grammar induction (WP5.3)

Aligners

- Align bilingual corpus (existing or output from WP4)
- Different levels of granularity
  - Sentence
  - Word
  - Chunk / Syntactic

Aligners

- Tools surveyed
  - Sentence
    - hunalign
  - Word
    - GIZA++, berkeleyaligner
    - word packing (“compound rich” languages, e.g. German)
  - Chunk
    - Marker hypothesis: Marclator
    - Syntactic: TreeAligner

Aligners

- Methodology
  - Integrate models: generative, syntactic, marker hypothesis
  - Extend range of language pairs
  - Tune to text type, domain and genre
  - Check/filter corpora acquired (comparability score)
  - Baseline: phrase alignment in Moses
  - Extrinsic evaluation (SMT in WP7)

Dictionary induction

Task: to derive bilingual dictionaries from aligned parallel corpus

Methodology

- Expectation-Maximisation algorithm
- Additional techniques on top of word correspondences → precision, fine-cleaning → reduce human intervention
- Go beyond word level: MW translations (NPs, MWEs)
- Baseline: word alignment in Moses
- Evaluation?

Transfer Grammars: Task

- Find criteria for lexical transfer selection
  - *not* meant:
    - structural transfer (Probst, Sánchez-Martínez, et al.)
      - (matching of POS-sequences
      - *independent* of lexical material)
    - bilingual term extraction (Cabré 2001, Gamallo 2007)
      - (does not care of 1:n situations)

- Classification:
  - structural transfer
  - lexical transfer
    - simple lexical
    - contextual lexical <- this is the task! *conditions* for transfer selection

Selection means used by current MT systems

- Word tagging
  - with domain / subject area information („MEDICAL“)
  - with locale / variant („EN_UK“ „DE_CH“)
- Morphosyntactic context
  - use information on local nodes (gender, number)
  - use structural contexts (arguments, prepositions, subcategorisation frames & fillers) (main means of RMT)
- Conceptual context
  - use conceptual environment for disambiguation
    - using word sense disambiguation, statistical word alignment

Focus of WP 5.3

- supervised learning of most important disambiguation means:
  1. domain tag assignment
  2. morphosyntactic tests
     - local features on gender / number
     - subcategorisation: Prepositions (for nouns and verbs)
     - presence / absence of verb arguments (trans./intrans.)
     - (relational Adj <-> compound specifier)
  1. conceptual contexts
     - *source* language concept clusters (SMT uses *target* language models)

Approach

- Preparation
  - Selection of disambiguation candidates (N, V, A)
  - Creation of parallel corpora
  - Creation of subcorpora for each translation
- Analysis and comparison
  1. domain tags: do subcorpora differ in domain?
  2. morphosyntactic:
     - gender: do they differ in gender? in number?
     - arguments: do they differ in transitivity? in subcategorised prepositions?
     - …
  1. conceptual: Can different SL concept clusters be built to disambigute?
- Verification with additional candidates or data

WP 5.3 Tools needed

- Standard pre-processing chain
  - Sentence Segmentiser, Tokeniser, Dictionary Lookup
- Analysis of transfer selection
  1. Domain tag assignment:
     - Topic classifier
  2. Morphosyntactic tests:
     - Parser to extract annotated subtrees
     - Tree matching component
  3. Conceptual context:
     - target-sensitive word sense disambiguation
- (Analysis of transfer actions
  - similar for the target side …) (if time permits)

Parallel corpora requirements

Quality:

- a really parallel (not comparable) corpora aligned on **sentence** level
- translation quality of aligned sentence pairs is essential for MT output

Linguistic pre-processing:

- tokenized plain text (plain PB-SMT)
- POS tagging, lemmatization (factored PB-SMT, EBMT)
- constitutency and dependency parsing (syntax motivated PB-SMT)

Size:

- for a baseline system: at least 1M sentece pairs (~20M words)
- for domain adaptation: 20K-200K sentece pairs (~400K-4M words)

LR Survey: Parallel corpora

Corpus	domain	French	Spanish	Italian	German	Greek
EuroParl *	parliamentary	52	49	47	42	27
JRC Acquis *	law	39	39	36	32	37
News Commentary	news	2	2		2	
United Nations	UN	205	190			
English-French	parliamentary	672				
EMEA *	medicine	14	14	14	12	17
OpenSubtitles	subtitles	5	15	1/2	2	
MLCC *	parliamentary	1	1	1	1	1
ECI/MCI	technical	15	15	1/2	1	1/2
ILSP	mix					2
IULA	technical		1 1/2			

- numbers in millions of words from English to the target language
- in corpora denoted by * all language pairs available

LR Survey: Monolingual corpora

Corpus	domain	English	French	Spanish	Italian	German	Greek
News (WMT)	news	1,113	107	107		315	
Gigaword	mix	3,000					
WaCky	mix	2,000	1,600		2,000	1,700	
BNC	mix	100					
ILSP EL corpus	news						140

- numbers in millions of words
- monolingual parts of the parallel corpora also available

LR Survey: Results

- A number of standard monolingual and parallel corpora available for all languages pairs of sufficient size & quality
- Parliamentary proceedings and debates can be considered „general domain“
- Monolingual web-crawled corpora available for English, French, German, Italian (WaCky) – unspecified domain
- No web-crawled parallel data available at all (Resnik's Strand is only a list of URLs, but quite outdated) – no fallback strategy

LR Survey: Proposal

- **EuroParl** for baseline systems
  - 40M words per language
  - all project language pairs available
  - parliamentary proceedings and dabates
  - quite general domain suitable for adaptation
- **Evaluation data** to be selected as a subset from webcrawled in-domain data (including 500-2000 sentence pairs for *test set* and *dev test set*)
- Focus on translation from **English to other languages** (but not all of them?)

Workplan t4-t14

Official deadlines:

- t6 Report on parallel technolgy tools (D5.1)
- t14 Aligners integrated in the platform (D5.2)
- t14 First MT evaluation (D7.2)

Internal deadlines:

- t6 decision on MT language pairs and domains
- t9 baseline MT systems trained
- t12 resources to be included in the first evaluation produced (D4.3)
- t12-t14 the first evaluation

Questions?

Thanks for your attention!

SMT Domain Adaptation

Assumption: general and in-domain monolingual and parallel data available

Possible approaches:

- one system build from mixture of the data
- two systems and a domain classifier (for sentences)
- two systems and system combination based on their n-best output

WP3 Questions & Comments

- Distribution of webservices across partners?
- Software requirements for webservices?
- Hardware specifications (no HW budget)?
- Example webservice wrapper?

WP4 Questions & Comments

- Rich text format support?
- Duplicate document/sentence detection?
- Distribution of webservices?
  - TPC tools for one language on one site?

WP7 Questions & Comments

- MT tools integrated into the platform?
  - alignment OK
  - language modelling?
  - phrase table extraction?
  - Decoding?
  - tuning?

- Only extrinsic automatic evaluation feasible

WP8 Questions & Comments

- Only extrinsic (MT) evaluation feasible