Hybrid Example-Based – Rule-Based MT: Feeding Apertium with Bilingual Chunks

Felipe Sánchez-Martínez

Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

Universitat d'Alacant
Universidad de Alicante

Work done in collaboration with
Andy Way (DCU) and Mikel L. Forcada (UA)
at the Centre for Next Generation Localisation – DCU

8th July 2009

Outline

Motivation & goal

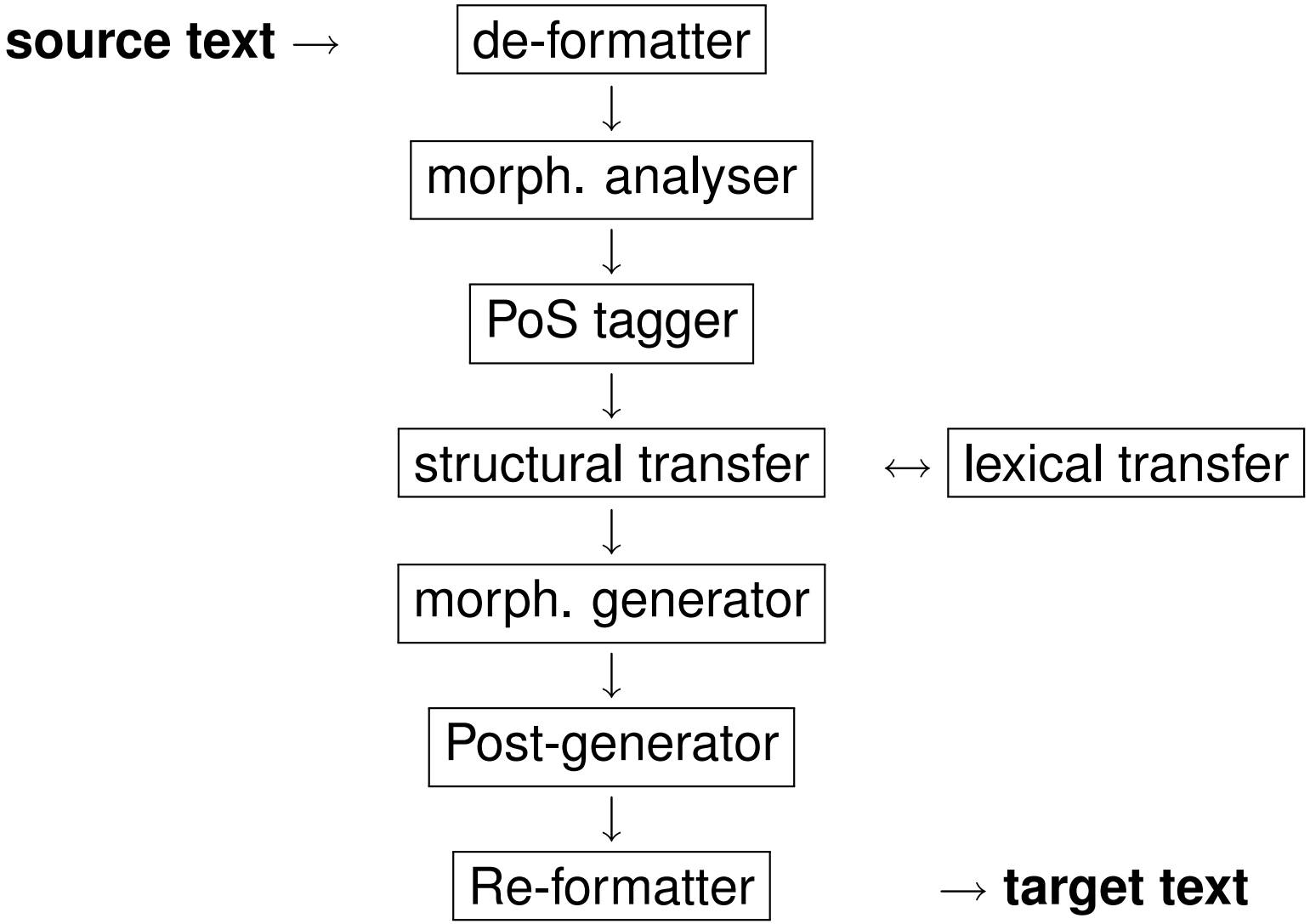
Motivation:

- Usually rule-based machine translation (RBMT) systems do not benefit from the post-edition effort of professional translators
- Some RBMT may benefit from the translation units found in translation memories (usually whole sentences)

Goal:

- To integrate sub-sentential translation units into the Apertium free/open-source MT platform
- Test the approach with bilingual chunks automatically obtained using the example-based methods implemented in Matrex

Apertium rule-based MT engine



Apertium: Example of execution /1

Source text:

Francis' car is broken

De-formatter:

Francis'[ ]car[ ]is broken

Morphological analyser:

^Francis'/Francis<np><ant><m><sg>+'s<gen>$[ ]
^car/car<n><sg>$[ ]^is/be<vbser><pri><p3><sg>$
^broken/break<vblex><pp>$

Part-of-speech tagger:

^Francis<np><ant><m><sg>$ ^'s<gen>$[ ]
^car<n><sg>$[ ]^be<vbser><pri><p3><sg>$
^break<vblex><pp>$

Apertium: Example of execution /2

Structural transfer (prechunk) + Lexical transfer:

^nom <SN><UNDET><m><sg>{^Francis<np><ant><3><4>$}$
^pr <GEN>{}$[ ]
^nom <SN><UNDET><m><sg>{^coche<n><3><4>$}$ [ ]
^be_pp <Vcop><vblex><pri><p3><sg><GD>{
^estar<vblex><3><4><5>$ ^romper<vblex><pp><6><5>$}$

Structural transfer (interchunk):

[]^nom <SN><PDET><m><sg>{^coche<n><3><4>$}$
[ ]^pr <PREP>{ ^de<pr>$}$
^nom <SN><PDET><m><sg>{ ^Francis<np><ant><3><4>$}$
^be_pp <Vcop><vblex><pri><p3><sg><m>{
^estar<vblex><3><4><5>$ ^romper<vblex><pp><6><5>$}$

Apertium: Example of execution /3

Structural transfer (postchunk):

[]^el<det><def><m><sg>$ ^coche<n><m><sg>$
[] ^de<pr>$ ^Francis<np><ant><m><sg>$
^estar<vblex><pri><p3><sg>$ ^romper<vblex><pp><m><sg>$

Morphological generator and post-generator:

[]el coche[ ]de Francis está roto

De-formatter:

el coche de Francis está roto

Target text:

el coche de Francis está roto

Considerations

To take into account:

- Not break the application of structural transfer rules
- Use the longest possible chunks

How can the application of rules be preserved?

- Introducing chunks delimiters as format information

  `... is [BCH_12_0]the chunk detected[ECH_12_0] by ...`
- Chunks can be then recognised after the translation

  `... es [BCH_12_0]el segmento detectado[ECH_12_0] por ...`

Known problem:

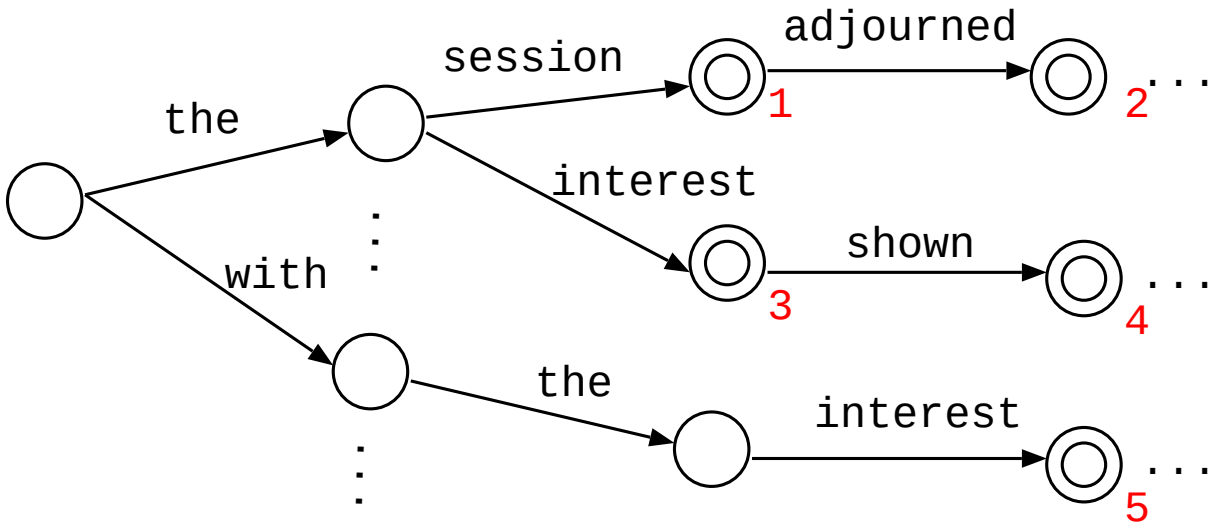
- As a result of the structural transfer rules, format information may be moved around
- Some rules also delete format information (known bug)

Translation approach

1. apply a dynamic-programming algorithm to compute the best coverage of the input sentence
2. translate the input sentence as usual by Apertium
3. use a language model to choose one of the possible translations for each of the bilingual chunks detected
   - One source-language chunk may have different target-language translations
   - Also consider Apertium translation

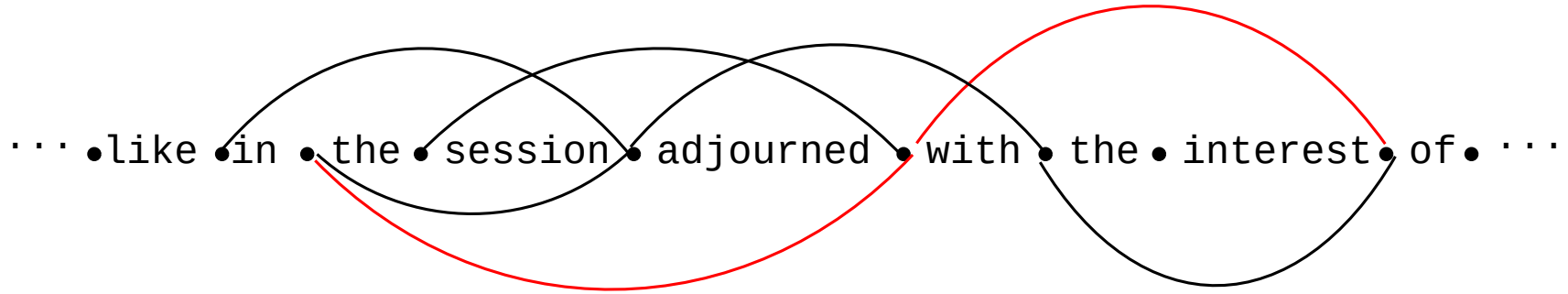
Computation of the best coverage: data structure

Store source-language chunks in a trie of strings



It allows to compute the best coverage in $O(l)$ time, where l is the length of the input sentence

Computation of the best coverage: algorithm



- A set of alive states in the trie is maintained to compute all the possible ways to cover the input sentence
- A new search is started at every word
- At each position the best coverage until that position is stored
- Is applied to text segments shorter than sentences
  - The best coverage can be retrieved when there are no more alive states

Computation of the best coverage

The best coverage:

- is the one that uses the least possible number of chunks
  - longest possible chunks
- each not covered word counts like one chunk
- if two coverages use the same number of chunks, the one that uses the most frequent chunks is used

Experimental setup /1

Data used:

- Corpora distributed for the WMT 09 Workshop for MT
- Language pairs: Spanish–English (`es-en`), English–Spanish (`en-es`)
- Linguistic data: `apertium-en-es`; SVN revision 9284

Software used:

- Apertium
- Giza++ and Moses to calculate word alignments and lexical probabilities
- SRILM to train 5-gram language models
- Matrex to segment training corpora and to align chunks

Experimental setup /2

Training corpus:

Max. sentence length: 45 words
Max. word ratio: 1.5 words (mean ration + std. dev.)

- # sent: 1,187,905; # `en` words: 26,983,025; # `es` words: 27,951,388

Development corpus:

- # sent: 2,050; # `en` words: 49,884; # `es` words: 52,719

Test corpus:

- # sent: 3,027; # `en` words: 77,438; # `es` words: 80,580

Experimental setup /3

Methods used to extract bilingual chunks:

- Marker-based bilingual chunks (using Matrex)
- Parse-tree based bilingual chunks (thanks to John Tinsley)
  - Preliminary results using previously compute chunks using an old version of the Europarl parallel corpus

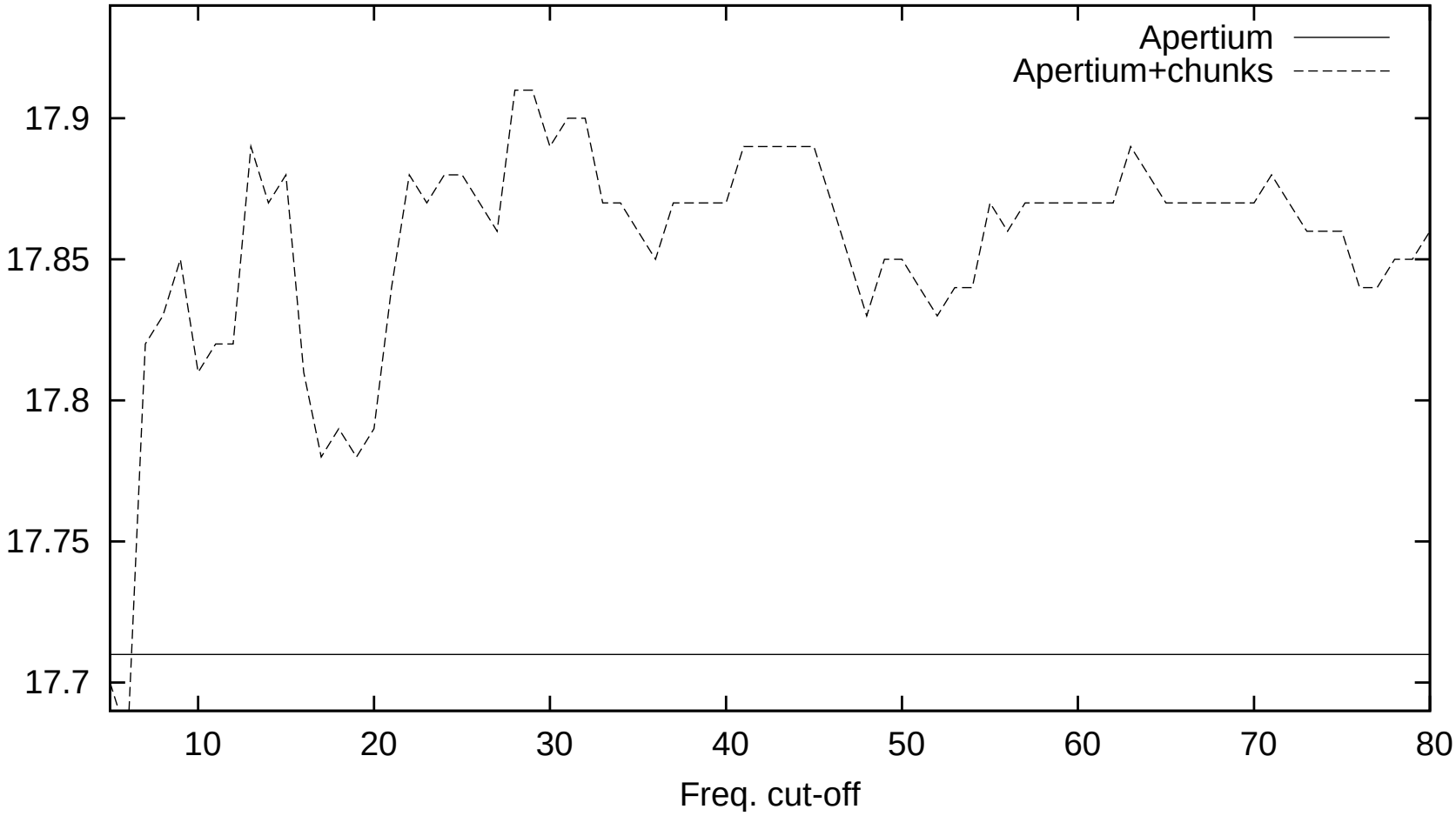
Results: marker-based chunks

Bilingual chunks filtering:

- There must be at least one word aligned in each side
- Chunks not seen at least N times are discarded
  - Tested values for N: $5 \ldots 80$
- Chunks containing punctuation marks and numbers are discarded

Results: marker-based chunks — Spanish→English /1

Development corpus (BLEU):



Results: marker-based chunks — Spanish→English /2

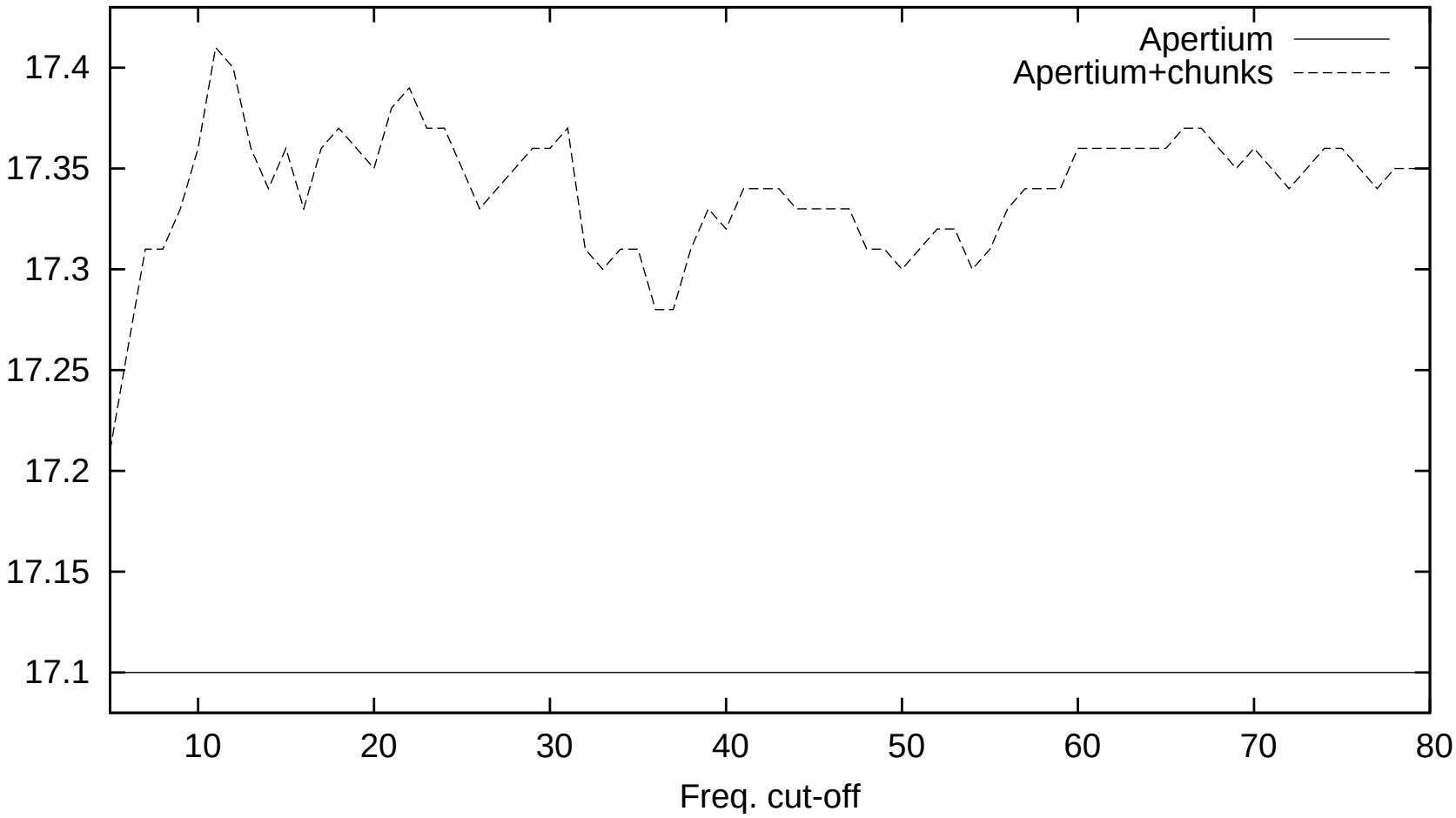
Test corpus (BLEU):

Apertium+chunks:	19.14
Apertium:	18.81

# of chunks:	6,600	(Freq. cut-off: 28)
# of applications:	6,321	
Words covered by chunks:	≈ 17%	
# of no Apertium:	2,559	(40%)
Words really covered by chunks:	≈ 6%	

Results: marker-based chunks — English→Spanish /1

Development corpus (BLEU):



Results: marker-based chunks — English→Spanish /2

Test corpus (BLEU):

Apertium+chunks:	18.94
Apertium:	18.51

# of chunks:	16,395	(Freq. cut-off: 11)
# of applications:	6,812	
Words covered by chunks:	≈ 18%	
# of no Apertium:	2,884	(42%)
Words really covered by chunks:	≈ 8 %	

Some Spanish→English examples /1

S:	desde hace muchos años un fenómeno misterioso ...
R:	for years , a mysterious phenomenon ...
A:	from does a lot of years a mysterious phenomenon ...
A+C:	for many years a mysterious phenomenon ...
S:	olmert devolvería casi todas las zonas ocupadas a cambio de la paz
R:	olmert would return ... territories in exchange for peace
A:	olmert it would give back ... zones to change of the peace
A+C:	olmert it would give back ... zones in exchange for peace
S:	pero hay una cosa que nos une :
R:	but there is one thing that connects us :
A:	but there is a thing that joins us :
A+C:	but there is one thing that joins us :

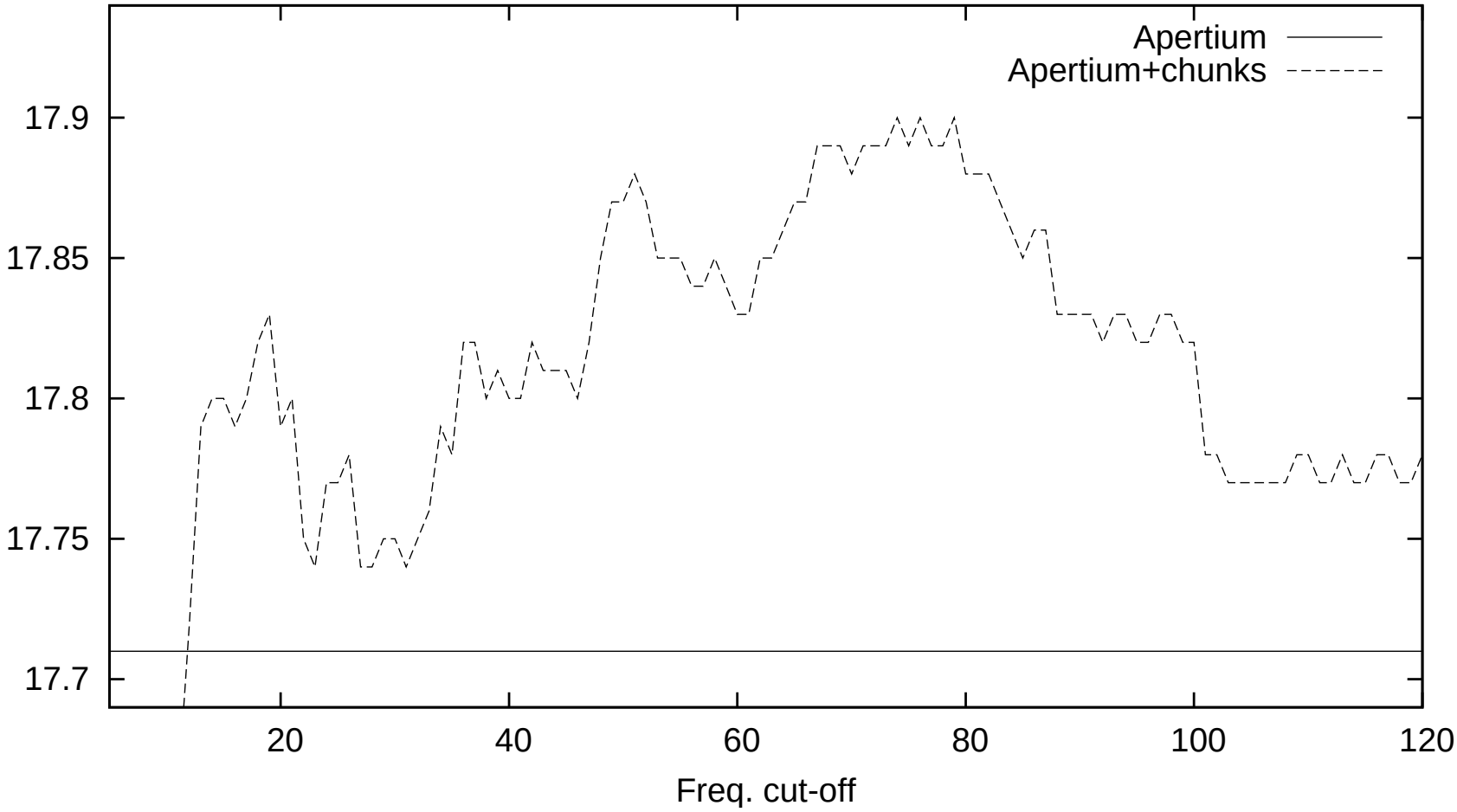
Results: tree-based chunks

Bilingual chunks filtering:

- There must be at least one target word aligned with a source word with $p(\text{target}|\text{source}) > 0.01$
- Chunks not seen at least N times are discarded
  - Tested values for N: $5 \ldots 120$
- Chunks containing punctuation marks and numbers are discarded

Results: tree-based chunks — Spanish→English /1

Development corpus (BLEU):



Results: tree-based chunks — Spanish→English /2

Test corpus (BLEU):

Apertium+chunks:	0.1911
Apertium:	0.1881

# of chunks:	7,466	(Freq. cut-off: 74)
# of applications:	4,650	
Words covered by chunks:	≈ 12%	
# of no Apertium:	3,075	(66%)
Words really covered by chunks:	≈ 4%	

Discussion /1

- Small improvement in both the development set and the test set
  - Better improvement in the test set

marker-based chunks	set	improvement	covered words
es-en	dev	+ 0.20	18% (7%)
	test	+ 0.33	17% (6%)
en-es	dev	+ 0.31	17% (7%)
	test	+ 0.43	18% (8%)

- Noise introduced due to how Apertium manages format information
  - Some chunks are not applied because chunk delimiters are lost
  - Some chunk delimiters are moved and the detected sequence of words after translation is not correct

Discussion /2

Possible way of improvement when computing the best coverage and two coverages uses the same number of chunks:

- Use the bilingual chunk that would produce the most-likely TL translation instead of the most frequent one
- How? Using a language model with gaps

◯ in the session ◯ ◯ with the interest .

Discussion /3

Future work:

- Try tree-based chunks obtained from the WMT 09 corpus using FreeLing to parse Spanish (in both directions)
- Perform a manual evaluation
- Combine both Spanish→English and English→Spanish chunks
  - Intersection
  - Union
- Try to know how Matrex is helping Apertium

Hybrid Example-Based – Rule-Based MT: Feeding Apertium with Bilingual Chunks

Felipe Sánchez-Martínez

Dept. Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

Universitat d'Alacant
Universidad de Alicante

Work done in collaboration with
Andy Way (DCU) and Mikel L. Forcada (UA)
at the Centre for Next Generation Localisation – DCU

8th July 2009