FBK's Machine Translation Systems for IWSLT 2012's TED Lectures

Nick Ruiz, Arianna Bisazza

Roldano Cattoni, Marcello Federico

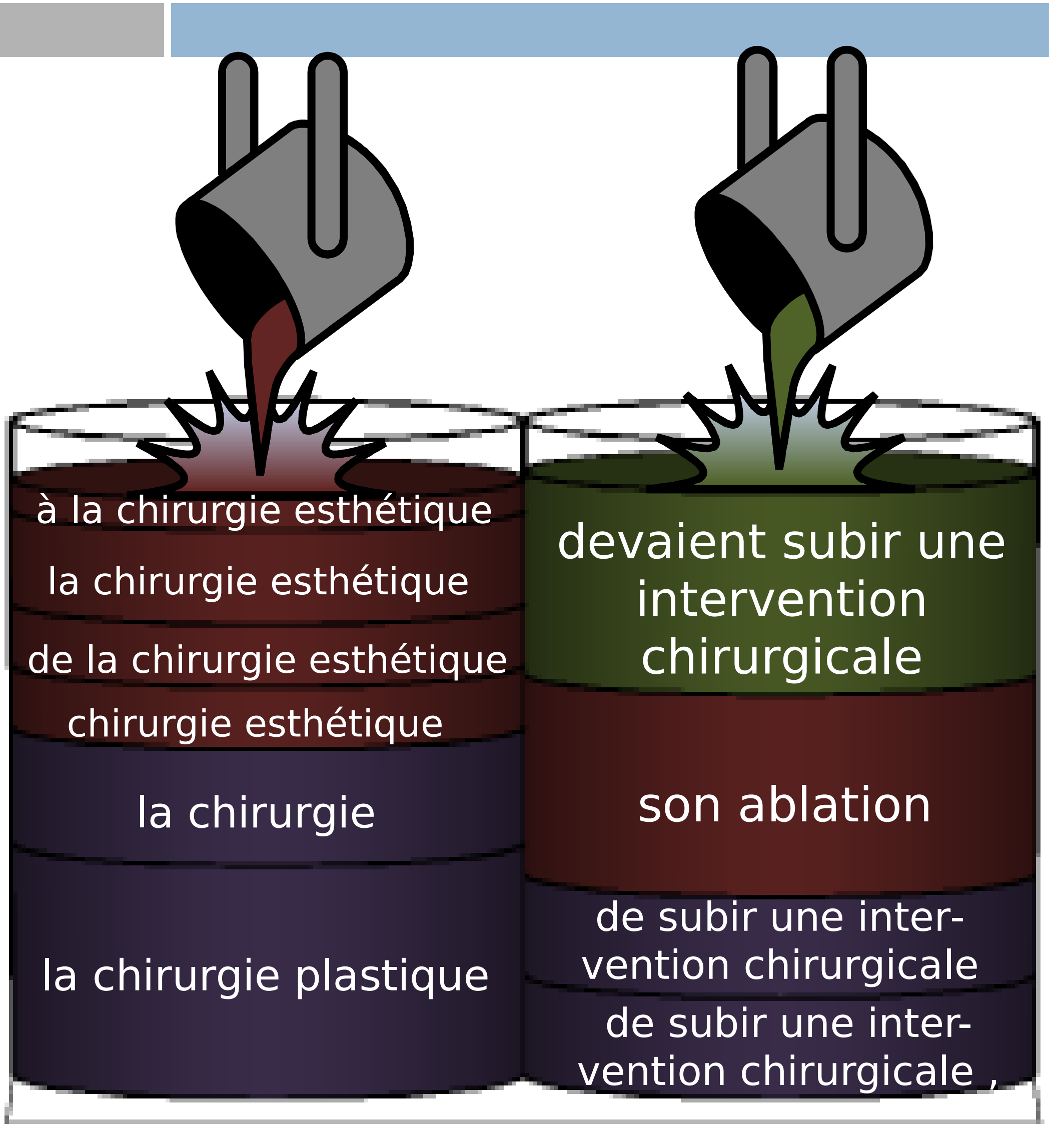
Hong Kong, 6 December 2012

Outline

- Common components
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion

Fill-Up

(Bisazza et al., 2011; Nakov, 2008)


cosmetic surgery
à la chirurgie esthétique
la chirurgie esthétique
de la chirurgie esthétique
chirurgie esthétique
la chirurgie
la chirurgie plastique
devaient subir une intervention chirurgicale
son ablation
de subir une inter-vention chirurgicale
de subir une inter-vention chirurgicale ,
to undergo surgery

Phrase Table
🟩 Europarl
🟥 Giga French
🟪 TED

Cross-Entropy LM Filtering

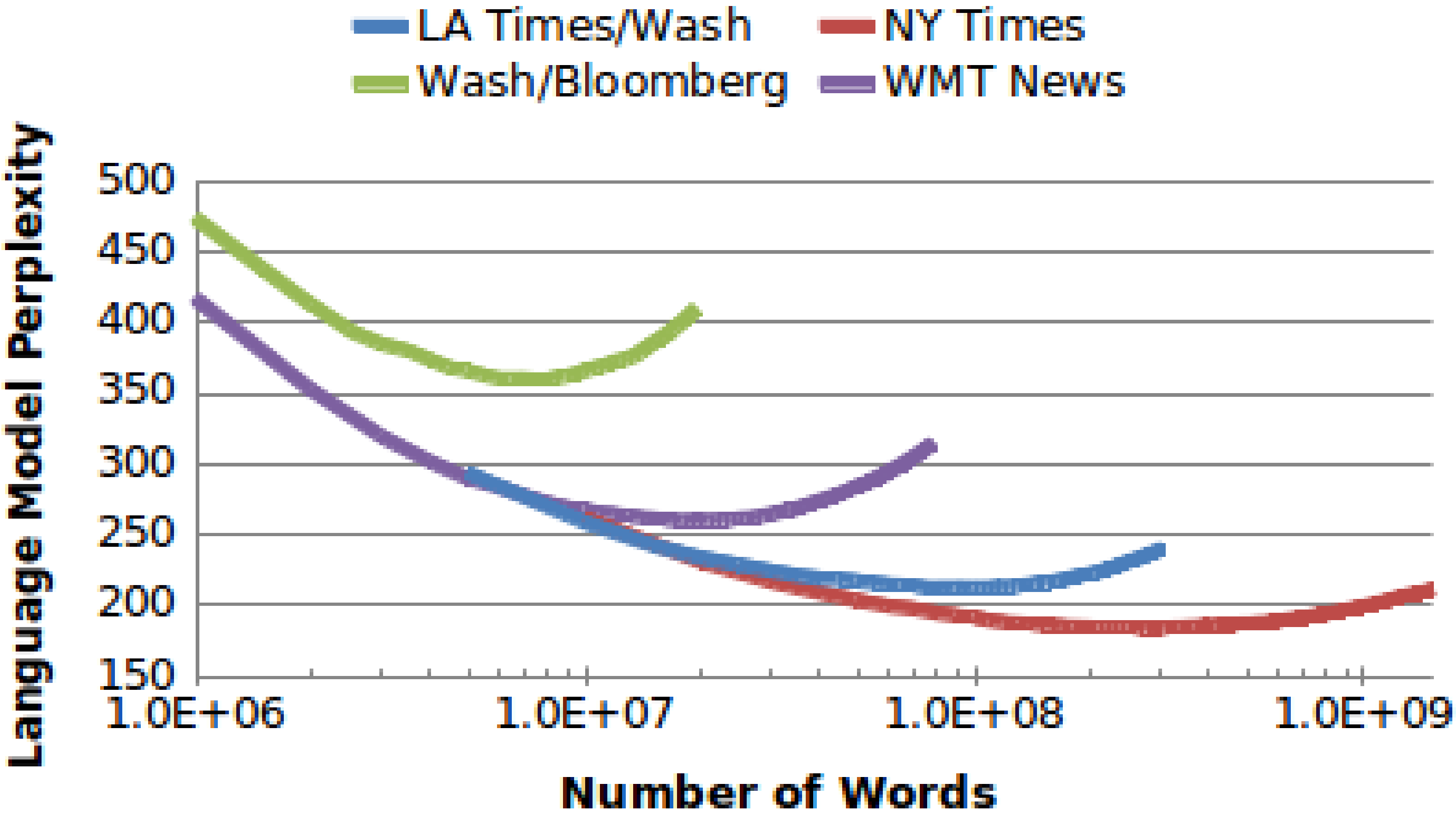
(Moore & Lewis, 2010)

- Cross-Entropy ranking of sentences in a out-of-domain corpus against TED
- Incrementally add sentences to minimize perplexity on a development set
- Also applicable to parallel corpora by filtering on target language

Cross-Entropy LM Filtering

(Moore & Lewis, 2010)

Cross-Entropy Filtering on English Corpora



Filtering tuned on TED dev2010 data

Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion

Arabic-English

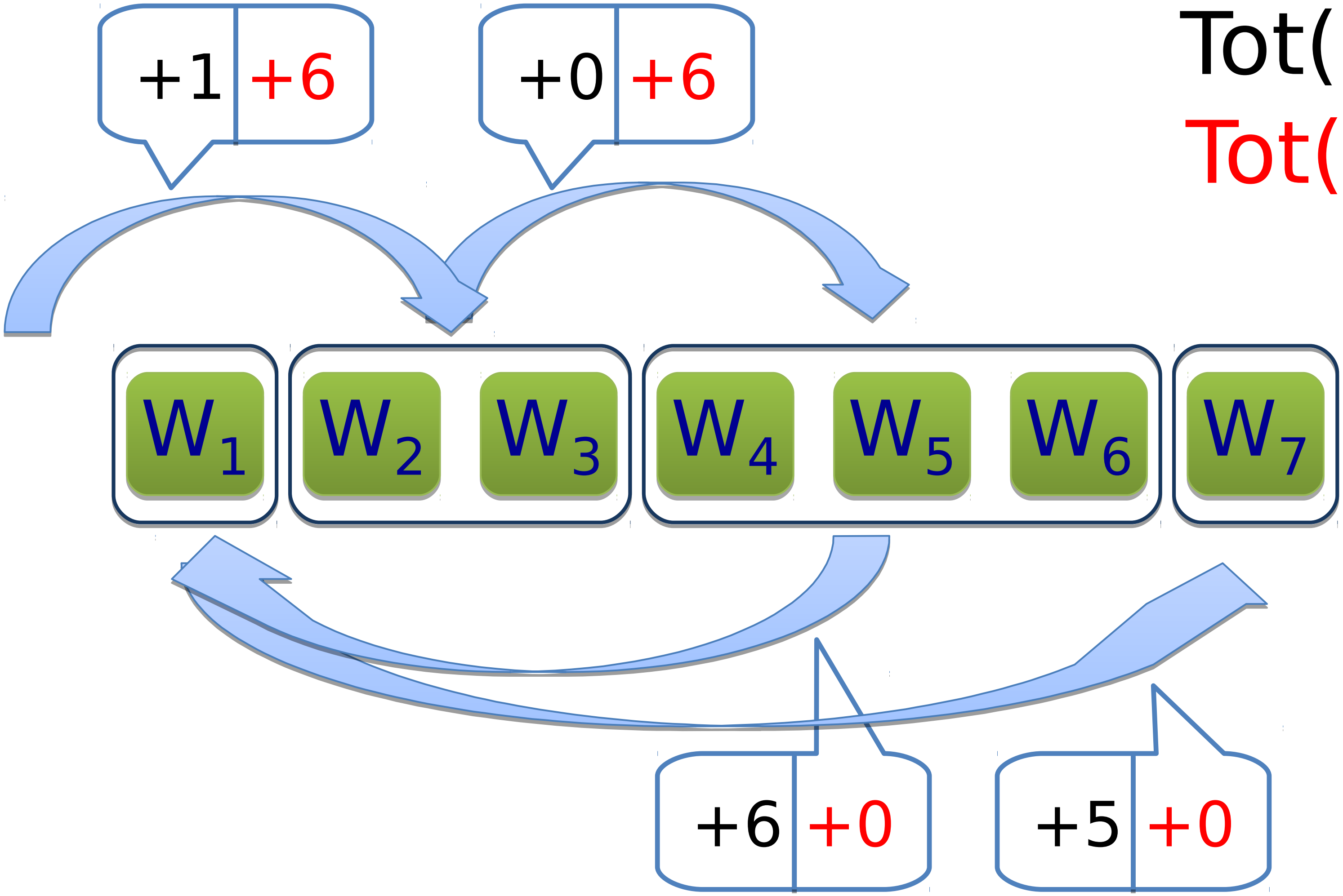
- Early Distortion Cost
- Hybrid Language Modeling
- Phrase/Reordering Fill-Up (TED+MultiUN)
- Mixture LM (TED, Gigaword, WMT News)

Early Distortion Cost

(Moore & Quirk, 2007)

- Improved distortion penalty
- Anticipates gradual accumulation of total distortion cost
  - Incorporates an estimate of future jump's cost
  - Same distortion penalty as standard distortion cost over a complete hypothesis
- Benefits: Improves comparability of translation hypotheses with the same number of covered words

Early Distortion Cost
(Moore & Quirk, 2007)

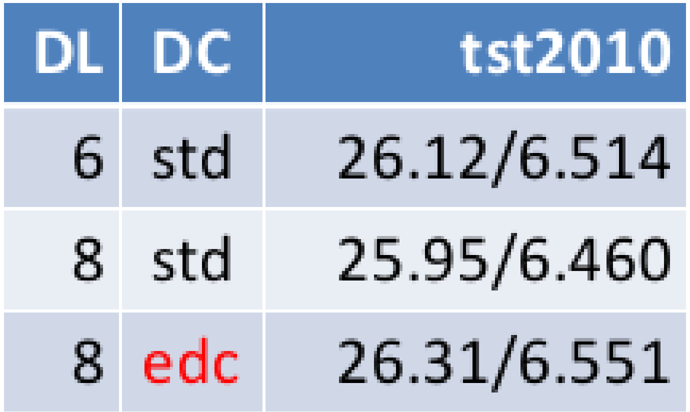


Tot(std) =12

Tot(edc)=12

Early Distortion Cost

(Moore & Quirk, 2007)



DL	DC	tst2010
6	std	26.12/6.514
8	std	25.95/6.460
8	edc	26.31/6.551

Hybrid Language Modeling

(Bisazza & Federico, 2011)

- Replace bottom 25% of tokens with POS tags – corresponds to 2% of types

In-domain target data

Now you **laugh**, but that **quote** has kind of a **sting** to it, right. And I think the reason it has…

Now you **VB** , but that **NN** has kind of a **NN** to it, right. And I think the reason it has…

…a sting is because thousands of years of history don 't reverse themselves without a lot of pain.

…a **NN** is because **NNS** of years of history don 't **VB** **PP** without a lot of **NN** .

Hybridly mapped word/POS data

- Allows for the construction of 10-gram LMs

Arabic-English results

	LM	DL	tst2011	tst2012
P	MixAll.4g +TED.Hybrid10g	8 edc	25.46/6.232	27.86/6.881
C_1	MixAll.4g	8 edc	25.19/6.205	27.74/6.903
C_2	MixFiltered.5g +TED.Hybrid10g	8 edc	25.13/6.190	27.54/6.828

Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion

Turkish-English

- Morphological Segmentation
- Hierarchical phrase-based decoding
- Mixture LM

Morphological Splitting

- Rule-based vs. Unsupervised segmentation

Distortion Limit	Distortion Calc	Seg	tst2010
15	std	MS6	13.61/5.280
15	std	MS15	14.38/5.273
15	std	Morfessor	13.45/5.080

- MS6: Nominal suffixes (case + possessive) only
- MS15: Nominal and verbal suffixes
  - e.g. person-subject, negation, passive, etc.
- Morfessor:
  - Concatenates non-initial “morphs” into word endings
  - Could perhaps be trained with better configurations

Morphological Splitting

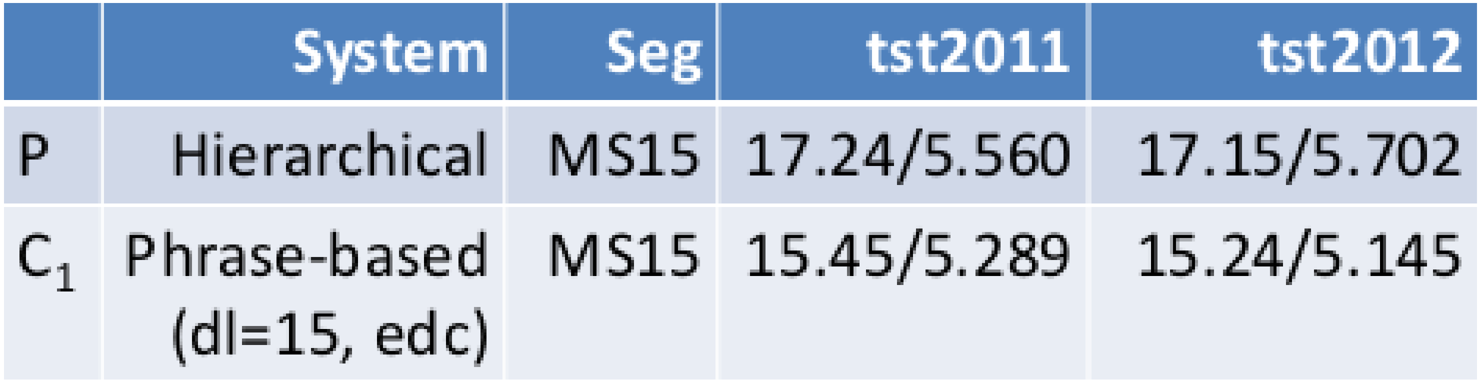
Original	Kendisine Don diyelim .					
Analyzed	**kendi**+Pron+Reflex +A3sg+P3sg+Dat		**don**+Noun+A3sg +Pnon+Nom	**de**+Verb+Pos +Opt+A1pl		.
MS15	**kendi**+Pron +Reflex+A3sg	+Dat	**don**+Noun+A3sg	**de**+Verb +Opt	+A1pl	.
Morfessor	Kendi	+sine	Don	diyelim		.
Trans	Let 's call him Don .					

Hierarchical Phrase-Based Decoding

- Better able to handle mismatches in predicate-argument structure between languages
- Robust with respect to long-distance reordering

Turkish (source)	English (target)	Rule
[X] söyle+Verb+Fut	will say [X]	SOV → SVO
[X] +Dat bak	look at [X]	S Comp V → S V Comp
[X] +Dat baktı	looked at [X]	S Comp V → S V Comp

Turkish-English results



	System	Seg	tst2011	tst2012
P	Hierarchical	MS15	17.24/5.560	17.15/5.702
C_1	Phrase-based (dl=15, edc)	MS15	15.45/5.289	15.24/5.145

Outline

- Common features
- Arabic-English
- Turkish-English
- Dutch-English
- Conclusion

Dutch-English

- Language properties
  - Similar to German
    - SVO for main clauses, SOV for subordinates
    - Noun casing, but less than German
  - Only “gendered” and “neutered” nouns/determiners
  - Compound nouns and verbs

Dutch-English

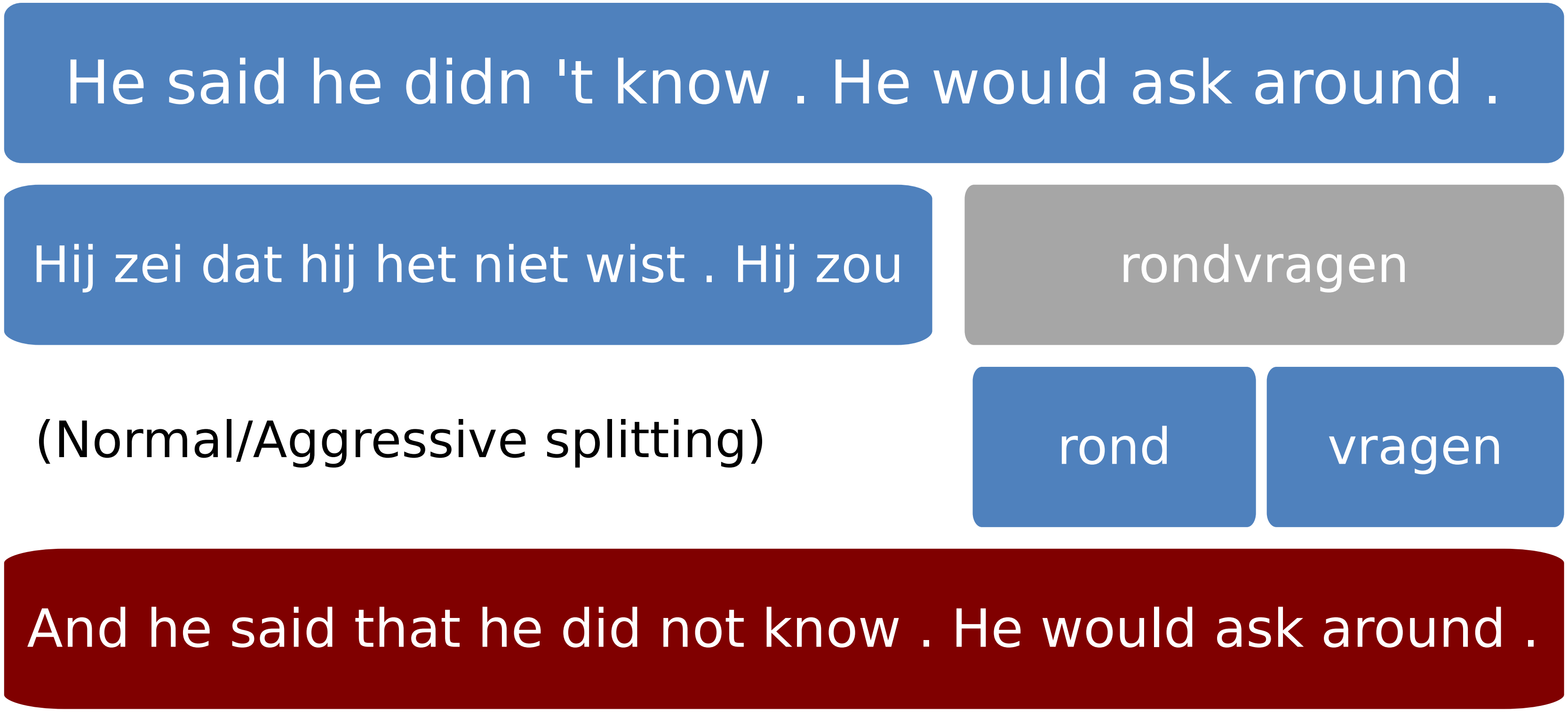
- Compound Splitting
- Phrase/Reordering Fill-Up (TED+Europarl)
- Mixture LM

Compound Splitting

(Koehn & Knight, 2003)

- Preliminary experiments on German, carried over to Dutch
- Moses Compound Splitting tool
  - Split candidate words into tokens already existing in a corpus' vocabulary
  - Default (normal) setting: min 4 characters per split
  - Aggressive setting: reduce minimum to 2 chars
    - e.g. “aanvragen”, “afvallen”

Compound Splitting



He said he didn 't know . He would ask around .

Hij zei dat hij het niet wist . Hij zou

rondvragen

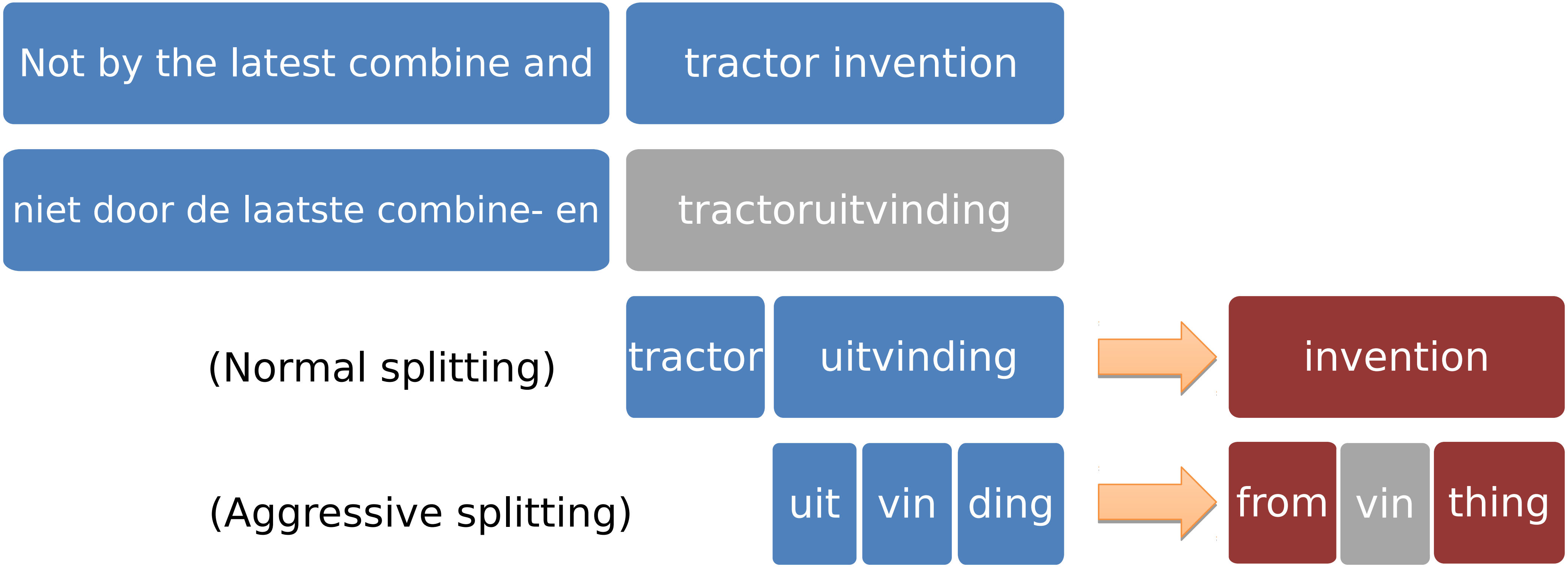
(Normal/Aggressive splitting)

rond

vragen

And he said that he did not know . He would ask around .

Compound Splitting



Not by the latest combine and

tractor invention

niet door de laatste combine- en

tractoruitvinding

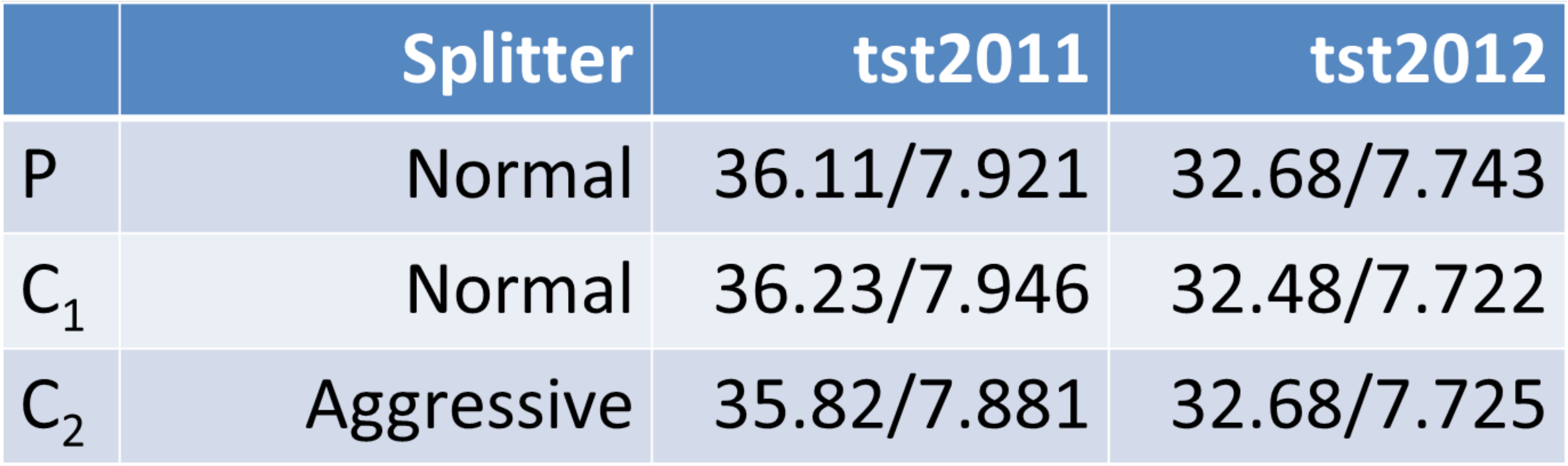
(Normal splitting)

tractor uitvinding → invention

(Aggressive splitting)

uit vin ding → from vin thing

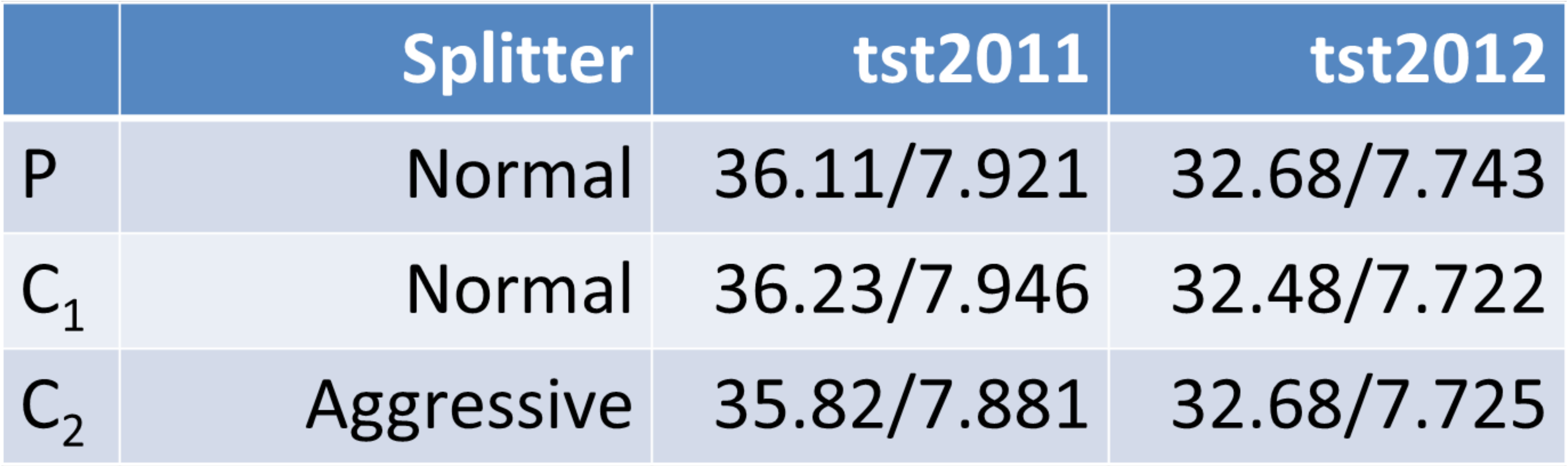
Dutch-English results



	Splitter	tst2011	tst2012
P	Normal	36.11/7.921	32.68/7.743
C_1	Normal	36.23/7.946	32.48/7.722
C_2	Aggressive	35.82/7.881	32.68/7.725

- P: 4-gram Mix LM
- C1: 5-gram Mix LM
- C2: 6-gram Mix LM

Dutch-English results



	Splitter	tst2011	tst2012
P	Normal	36.11/7.921	32.68/7.743
C_1	Normal	36.23/7.946	32.48/7.722
C_2	Aggressive	35.82/7.881	32.68/7.725

- P: 4-gram Mix LM
- C1: 5-gram Mix LM
- C2: 6-gram Mix LM

Conclusion

- We present several ideas for Arabic-, Turkish-, and Dutch-English machine translation
- Contributions:
  - Early distortion limit (Arabic, attempted w/ Turkish)
  - Morphological Segmentation (Turkish)
  - Compound Splitting (Dutch)
  - Corpora Filtering