qtleap

TOWARDS DEEPER MT – A HYBRID SYSTEM FOR GERMAN

ELEFTHERIOS AVRAMIDIS, MAJA POPOVIC,
ALJOSCHA BURCHARDT AND HANS USZKOREIT

DMTW 2015, SEPTEMBER 3-4, PRAGUE

THE CHALLENGE: PRECISION OR RECALL?

- Current statistical MT systems internally have a high recall in terms of the right translation bits being present somewhere in the search space
    - Ensuring precision in terms of the chosen/generated output being a good translation is difficult
- Deep (knowledge-driven, transfer-based) systems can have high precision (up to always correct)
    - Recall is a problem: parsing failure or gaps in the lexicon typically lead to a dead-end
    - Precision suffers from missing statistical evidence

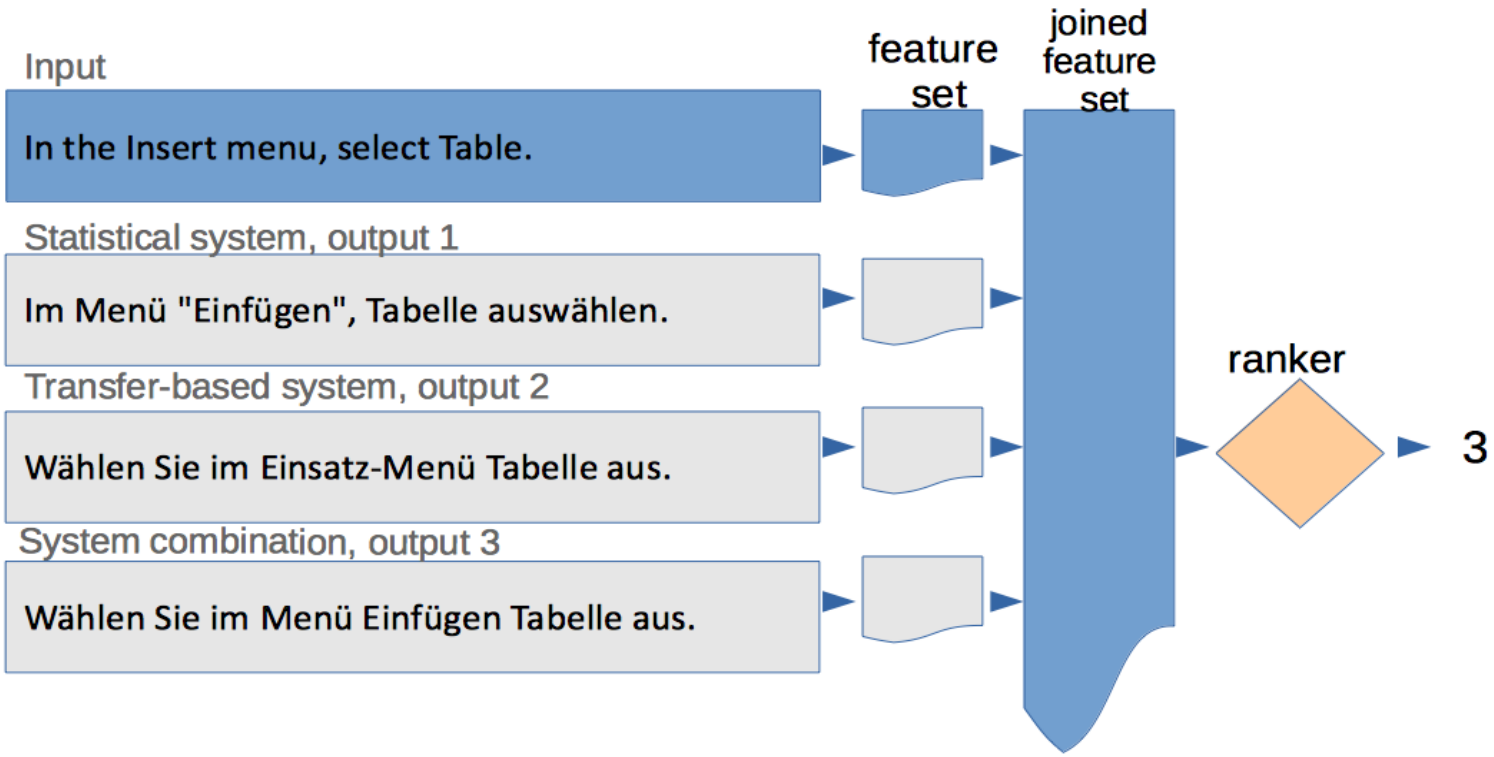
BASIC OPTIONS WHEN GETTING DEEPER

- Try to drastically improve recall (and also precision) of purely knowledge-driven systems
- Try to improve precision of statistical systems by using more linguistically informed pre-editing/models/selection/post-editing/ etc.
- Do both in a hybrid setting (the QTLeap way)

A HYBRID SYSTEM FOR EN<>DE

- System 1:
- A statistical Moses system,
- the commercial transfer-based system Lucy,
- their serial system combination,
- a linguistically informed selection mechanism (“ranker”).

HYBRID STRATEGY



Human reference: *Wählen Sie im Einfügen Menü die Tabelle aus*

YESTERDAY AT A METRO STATION

THE SYSTEMS IN A NUTSHELL

- Vanilla phrase-bases Moses trained on general domain and "technical help" domain (Libreoffice, Drupal, Ubuntu, etc.)
- Commercial Lucy RbMT performing analysis, transfer, and generation. A RestAPI allows the different processing steps and/or intermediate results to be influenced.
- Serial Transfer+SMT system combination.

SELECTION MECHANISM 1/3

- Automatic syntactic and dependency analysis is employed on a sentence level, in order to choose the sentence that fulfills the basic quality aspects of the translation:
    a) assert the fluency of the generated sentence, by analyzing the quality of its syntax
    b) ensure its adequacy, by comparing the structures of the source with the structures of the generated sentence.
- Ranker based on machine learning against training preference labels.

SELECTION MECHANISM 2/3

- Feature sets:
  - **Basic syntax-based feature set**: unknown words, count of tokens, count of alternative parse trees, count of verb phrases, parse log likelihood.
  - **Basic feature set + 17 QuEst baseline features**: this feature set combines the basic syntax-based feature set described above with the baseline feature set of the QuEst toolkit. This feature set combination obtained the best result in the WMT13 quality estimation task.
  - **Basic syntax-based feature set with Bit Parser**: here we replace the Berkeley parser features on the target side with Bit Parser.
  - **Advanced syntax-based feature set**: this augments the basic set by adding IBM model 1 probabilities, full depth of parse trees, depth of the 'S' node, position of the VP and other verb nodes from the beginning and end of the parent node, count of unpaired brackets and compound suggestions (for German, as indicated by LanguageTool.org).

SELECTION MECHANISM 3/3

- Best feature sets:
  - The *basic syntax-based feature set* for English-German, trained with Support Vector Machines against METEOR scores.
  - The *advanced syntax-based feature set* for German-English, trained with Linear Discriminant Analysis against METEOR scores.
- Selection on QTLeap corpus:

	Transfer	SMT	Transfer+SMT
de→en questions	45.2%	33.3%	23.8%
en→de answers	42.5 %	16.3%	50.5%

Table 1: Percentages chosen automatically by the selection mechanism from each of the systems. Percentages which sum more than 100% indicate ties. When ties occur, there is a preset order of preference SMT, Transfer, Transfer+SMT.

RESULTS ON QTLEAP CORPUS

		questions de→en	answers en→de
Moses	BLEU	43.0	41.7
	wordF	44.6	42.2
	charF	64.9	64.7
System 1	BLEU	43.3	33.0
	wordF	43.8	30.2
	charF	63.4	57.4

Table 2: BLEU scores, word-level and character-level F-scores for Moses baseline and System 1 translation outputs.

BREAKDOWN OF ERROR TYPES

		questions de→en	answers en→de
Moses	form	1.2	4.4
	order	6.5	5.7
	omission	4.4	4.6
	addition	2.8	3.7
	mistranslation	12.2	11.9
System 1	form	1.1	4.0
	order	5.6	5.6
	omission	3.4	3.0
	addition	3.6	7.4
	mistranslation	12.8	13.5

Table 3: Class error rates for Moses and System 1 translation outputs.

USER EVALUATION

- Compare Moses and System 1 (randomised of course):
  i. A is a better answer than B
  ii. B is a better answer than A
  iii. A and B are equally good answers
  iv. A and B are equally bad answers
- 100 question-answer pairs were judged by three volunteers. If we lump ties (i.e., iii and iv) together, the central (averaged) results of the user evaluation are:
  - **System 1** has been judged **better** than Moses in **17.3%** of cases (i)
  - **System 1** has been judged **better or same** as Moses in **75.5 %** of cases (i +iii+iv)

USER EVALUATION EXAMPLE

Example where System 1 wins:

Ref: Ja, können Sie. Beide Technologien sind kompatibel.

Moses: Ja, Sie können. Beide Technologien kompatibel sind.

Sys.1: Ja , Sie können. Beide Technologien sind zueinander passend.

WMT 2015 (FORTHCOMING) – OBSERVATIONS

qtleap

(a) De→En

German→English		Lucy	Moses	LucyMoses
ML classifier		42.1	36.6	21.3
POS 4-gram IBM1	L+M	2.8	97.2	/
	L+LM	2.5	/	97.5
	LM+M	/	42.4	57.6
	L+LM+M	1.7	56.0	42.3
WORDF	L+LM+M	29.3	31.8	38.9
POSF	L+LM+M	34.5	33.7	31.8

Upper bounds

(b) En→De

English→German		Lucy	Moses	LucyMoses
ML classifier		44.0	8.0	48.0
POS 4-gram IBM1	L+M	56.5	43.5	/
	L+LM	63.3	/	36.7
	LM+M	/	45.5	54.5
	L+LM+M	41.5	22.1	36.3
WORDF	L+LM+M	34.2	29.4	36.3
POSF	L+LM+M	42.3	27.1	30.5

Table 2: Percentage of selected sentences from each individual system.

WMT 2015 RESULTS

(a) De→En

German→English			BLEU	WORDF	POSF
individual systems		Lucy (L)	20.8	25.9	42.6
		Moses (M)	23.2	28.2	42.7
		LucyMoses (LM)	23.2	27.9	44.2
selection mechanism	ML classifier	L+LM+M	22.6	27.4	43.6
	POS 4-gram IBM1	L+M	23.2	28.2	42.8
		L+LM	23.2	27.9	44.2
		LM+M	23.7	28.6	44.5
		L+LM+M	23.7	28.6	44.5
upper bounds	max(WORDF)	L+LM+M	**26.9**	**30.8**	46.8
	max(POSF)	L+LM+M	25.6	30.7	**48.6**

(b) En→De

English→German			BLEU	WORDF	POSF
individual systems		Lucy (L)	17.3	22.9	44.5
		Moses (M)	17.1	23.1	41.9
		LucyMoses (LM)	18.9	24.4	45.3
selection mechanism	ML classifier	L+LM+M	18.1	23.7	44.4
	POS 4-gram IBM1	L+M	18.2	23.6	44.7
		L+LM	18.6	24.0	45.7
		LM+M	19.1	24.4	45.1
		L+LM+M	18.9	24.1	45.4
upper bounds	max(WORDF)	L+LM+M	**22.4**	**26.6**	47.1
	max(POSF)	L+LM+M	21.0	26.1	**49.4**

Table 1: Translation results [%] for the German-English language pair.

DIFFERENCES BETWEEN SELECTION RESULTS

4)	src:	Über mehrere Jahre hatte niemand in dem Haus gelebt.
	ref:	No one had lived in the house for several years.
WORDF, POSF	Lucy:	Over several years nobody had lived in the house.
IBM1	Moses:	No one had over several years lived in the House.
MLC	LucyMoses:	For several years, no one had lived in the House.

OUTLOOK

- Improvement on the lexical level (ongoing):
  - Special lexicons (Gazetteers)
  - WSD
  - Translation of items like „File > Save As“
  - Etc.
- Improvement on the structural level (future work):
  - Order of constituents (e.g., temporal phrases)
  - Long-distance phenomena (e.g., verb prefixes in German)
  - System combination on the phrasal level
  - Etc.
- Further evaluation and improvement of the selection mechanism

qtleap

TRANSFER-BASED SYSTEM