Using Word Posterior in Lattice Translation

Vicente Alabau

Institut Tecnològic d'Informàtica

e-mail: valabau@iti.upv.es

October 16, 2007

Index

- Motivation
- Word Posterior Probabilities
- Translation System
- Results
- Conclusions and Future Work

Motivation - Common approaches

- Serial approach:
  - + simple and fast - propagates errors from ASR
- Semi-coupled approach:
  - n-best: + simple - redundancy, time-consuming
  - *lattice*: + full searched space - time-consuming
  - *confusion network*: + simplified lattice, efficient - loss of grammar
- Integrated approach:
  - + theoretically promising - bad performance on non-simple corpora

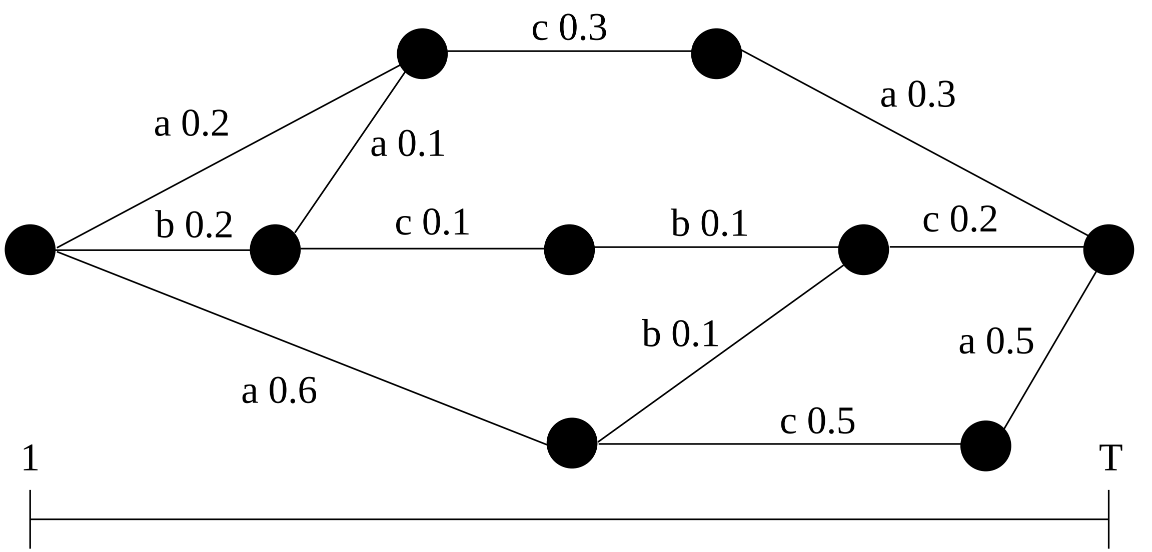
Word Posterior Probabilities

- Motivation
  - One should maximize word posterior probabilities to minimize WER (Mangu00)
  - Confusion networks (Bertoldi05):
    * word posterior probabilities
    * lattice simplification
- Our approach
  - Word posterior probabilities over a lattice
  - Take advantage of techniques in confidence measures (Sanchis04)

Word Posterior Probabilities: Forward-Backward

- being w the hypothesized word, s the start node and e the end node:

$$P([w,\, s,\, e] \mid \vec{x}_1^T) = \frac{1}{P(\vec{x}_1^T)} \sum_{\substack{f_1^J \in G : \\ \exists [w', s', e'] : \\ w' = w,\, s' = s,\, e' = e}} P(f_1^J,\, \vec{x}_1^T) \quad (1)$$

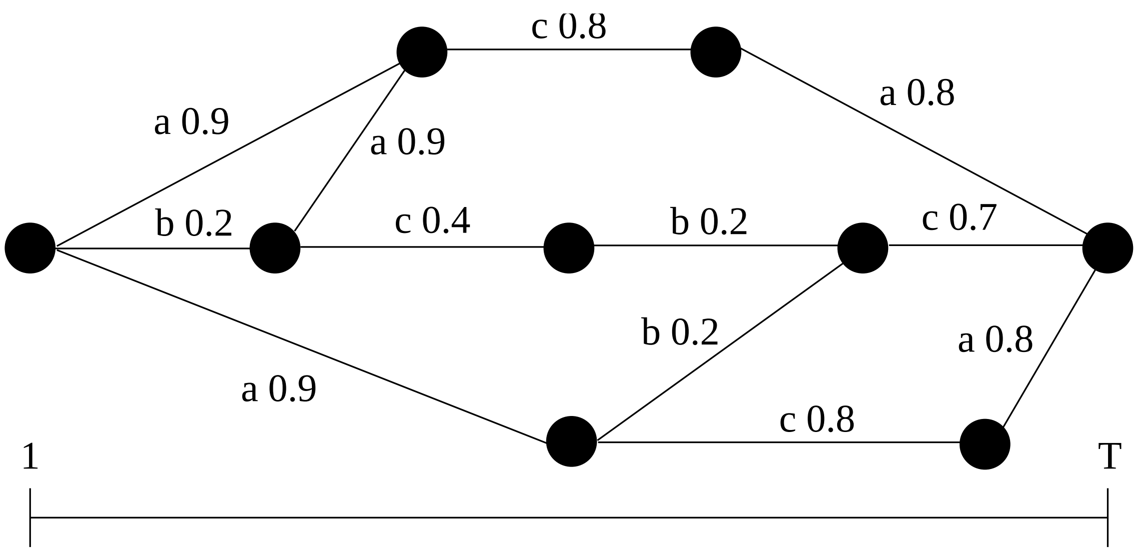


Word Posterior Probabilities

- maximum of the frame time posterior probability (Wessel01)

$$P_t(w \mid \vec{x}_1^T) = \sum_{t \in [s', e']} P([w,\, s',\, e'] \mid \vec{x}_1^T) \tag{2}$$

$$P([w,\, s,\, e] \mid \vec{x}_1^T) = \max_{s \leq t \leq e} P_t(w \mid \vec{x}_1^T) \tag{3}$$



Translation System

- Log-linear model:
  - Word posterior probabilities
  - GIATI:
    * Joint probability model
    * N-grams of bilingual pairs
    * 5-gram (w/o cutting off)
    * integrated lattice search
    * monotonous search
  - Output word penalty
  - Output language model (5-gram)

Translation System

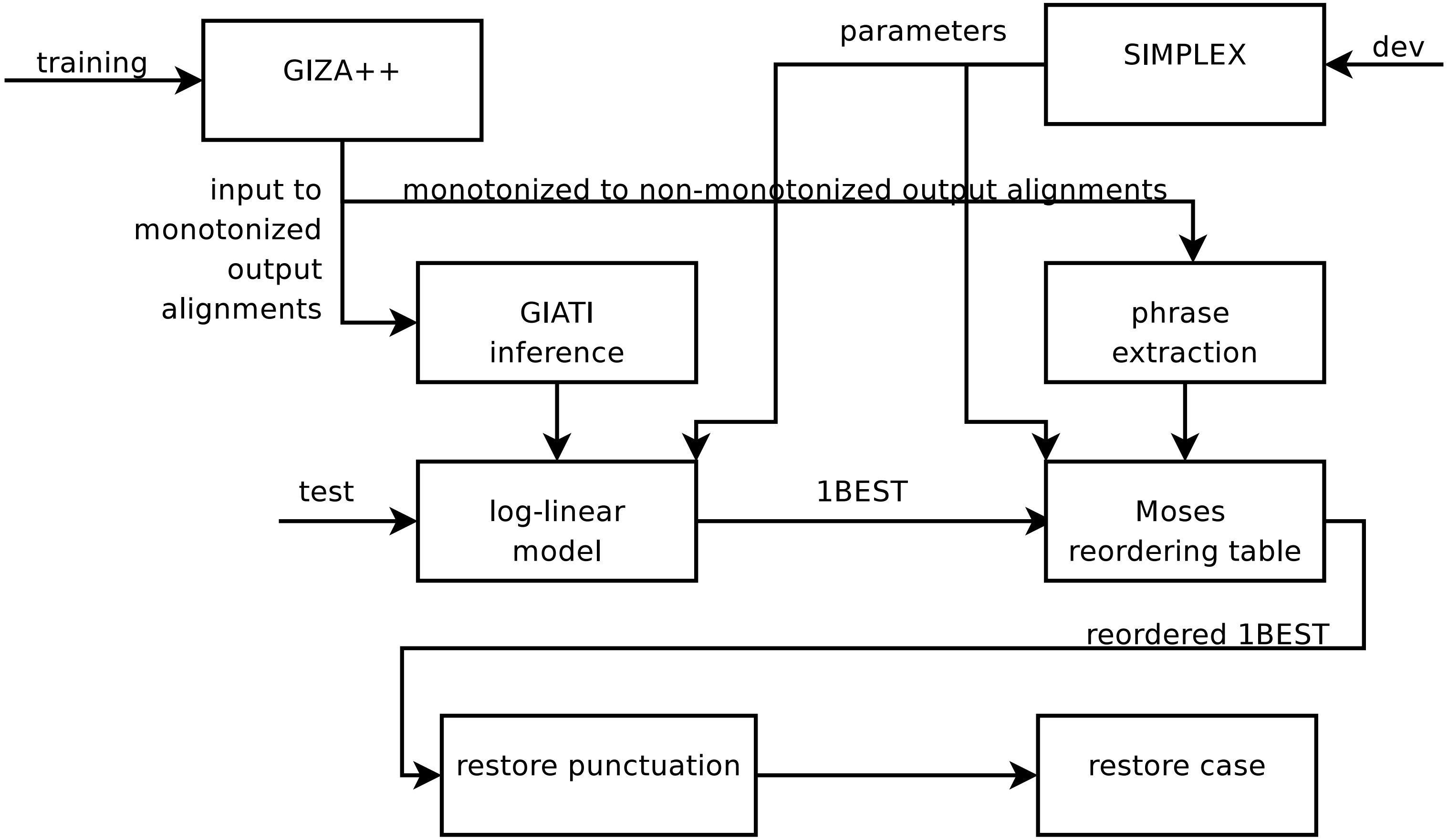
- Reordering:
  - Serial, 1BEST approach
  - Monotonization of the output
  - Translate with moses from monotonized to regular word order
  - Models: reordering table and output language model
  - Monotonous search

Preprocess and postprocess

- Preprocess:
  - Case and punctuation were removed from training
  - Sentence splitting at sentence boundaries (.?!)
  - Lattice pruning

- Postprocess:
  - Punctuation and case restoration: IWSLT06 method using SRILM
  - Capitalization after punctuation marks

System architecture



Corpus statistics

		Italian	English
Train	Sentences	19971	
	Running words	$172k$	$189k$
	Vocabulary	$10,152$	$7,165$
Dev4	Sentences	489	
	Running words	$4,831$	$6,848$
	OOV words	224	208
Dev5a	Sentences	500	
	Running words	$5,607$	$7,491$
	OOV words	296	264
Dev5b	Sentences	996	
	Running words	$8,487$	$11,968$
	OOV words	591	611
Test	Sentences	724	
	Running words	$6,420$	$9,054$
	OOV words	542	439

Effect of adding features to the baseline model

- Primary run: 16.13 BLEU

	dev4		dev5a		dev5b		test	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
baseline	36.29	7.59	31.96	7.06	12.53	4.02	22.80	5.49
+WP	37.45	7.35	32.55	6.82	14.07	3.77	19.56	5.06
+OL	37.06	7.42	32.55	6.91	12.37	3.82	22.32	5.25
+WP+OL	38.19	7.20	32.67	6.66	13.44	4.20	21.83	5.57
+RM	37.53	7.95	32.74	7.41	13.94	4.30	23.92	5.79
+WP+OL+RM	38.98	7.81	32.86	7.18	14.34	4.37	23.22	5.86

- *WP,* output word insertion penalty
- *OL,* output language model
- *RM,* reordering model

Effect of adding dev corpus to the training corpus

- Primary run: 16.13 BLEU

	w/o dev		with dev	
	BLEU	NIST	BLEU	NIST
baseline	22.80	5.49	31.29	6.66
+WP	22.09	5.56	12.16	2.97
+OL	22.79	5.52	30.83	6.64
+WP+OL	21.79	5.56	11.89	2.91
+RM	23.46	5.74	32.28	6.95
+WP+OL+RM	23.22	5.86	31.21	6.77

- *WP,* output word insertion penalty
- *OL,* output language model
- *RM,* reordering model

Results for different input conditions

	dev4		dev5a		dev5b		test	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
1BEST	33.53	6.92	26.97	6.12	13.21	4.19	21.50	5.56
LAT	33.69	6.95	27.24	6.14	13.35	4.16	18.71	5.22
GER	34.11	7.02	27.49	6.18	13.90	4.29	22.64	5.77
CLEAN	38.98	7.81	32.86	7.18	14.34	4.37	23.22	5.86

- *LAT,* lattice with word posterior probabilities
- *GER,* using the sentence from the lattice with less word error rate

Conclusions

- Word Posterior approach
  - Results not conclusive
  - Small differences between 1BEST and CLEAN scores
  - Some improvements were achieved
  - Needs work on pruning
- Adding devset to training matters

Future Work

- Comparison with n-best, confidence measures, lattice with acoustic scores
- Add additional state-of-the-art confidence features
- Add translation features
- Features based on multiple lattices
- Lattice reduction

Thank you for your attention!

Vicente Alabau

valabau@dsic.upv.es

References

[Mangu et al., 2000] Mangu, L., Brill E., and Stolcke A. (2000) Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks, In *Computer, Speech and Language*, 14(4):373-400.

[Wessel et al., 2001] Wessel, F., Schluter, R., Macherey, K., and Ney, H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, 9(3).

[Bertoldi and Federico, 2005] Bertoldi, N. and Federico, M. (2005). A new decoder for spoken language translation based on confusion networks. In *IEEE Automatic Speech Recognition and Understanding Workshop*.

[Sanchis, 2004] Sanchis-Navarro, J.A. (2004) Estimación y aplicación de medidas de confianza en reconocimiento automático del habla. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia Tesis Doctoral en Informática