Workpackage 2
Interactive Translation Prediction
2nd review meeting

V. Alabau, G. Sanchis-Trilles, F. Casacuberta, Universitat Politècnica de València
P. Koehn, University of Edinburgh

November 25, 2013

Work Package 2: Contents

- Goal: Render interaction process flexible and efficient
- Highly active during the second year
- Basic research involving:
  - Task 2.1. Months 1-18. Novel prediction strategies
  - Task 2.2. Months 6-24. Multi-modality
  - Task 2.3. Months 13-36. Use of confidence measures
  - Task 2.4. Months 6-18. Introduction of syntax-based translation models

Index

Introduction

- Despite important advances in SMT, human supervision required for ensuring quality
- Interactive translation prediction (ITP) paradigm:
  - Combine MT systems with human expert knowledge
    - MT system provides efficiency
    - Human expert provides quality
  - User feedback in form of corrections in the translation
  - ITP provides best completion for a partially validated sentence
  - Assumption: translation process is left-to-right

    $\boldsymbol{x}$: sentence in source language

    $\boldsymbol{y}$: sentence in target language

    $\boldsymbol{p}$: validated prefix

    k : current word typed (correction)

    $\boldsymbol{s}$ suffix generated (prediction)

    $(\boldsymbol{p}k\boldsymbol{s}) = \boldsymbol{y}$

ITP Example

SOURCE (x): Para imprimir una lista de fuentes postscript:
REFERENCE (y): To print a list of postscript fonts:

ITER-0	(p)	
ITER-1	$(\hat{s})$	*To print a postscript font list:*
	(p)	To print a
	(k)	list
FINAL	$(\hat{s})$	*of postscript fonts:*
	(k)	$(\#)$
	$(p \equiv y)$	To print a list of postscript fonts:

$\Rightarrow$ Sentence corrected with just one word-stroke (vs. four in post-edition)
$\Rightarrow$ Interaction also possible at the character level

- Leverage fully-fledged SMT systems for ITP:

$$\hat{s} = \underset{s}{\text{argmax}}\, p(s \mid x, p, k)$$

with $p \equiv$ validated prefix, $k \equiv$ current word typed, $s \equiv$ suffix generated, $(pks) = y$

Task 2.1

Search and Machine Learning Criteria

Task 2.1: Search and machine learning criteria

- Schedule: months 1–18
- Goal: basic research attempting to improve prediction as such
- Approaches:
  - Research novel error recovery strategies
  - Re-consider the prediction problem from a theoretical point of view:
    * optimum decision rule
    * word prediction as a machine learning problem
- Focus: improve efficiency and accuracy of ITP

Optimum Decision Rule for ITP: Motivation

- Classical suffix search (as in auto-completion):

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s} \mid \boldsymbol{x}, \boldsymbol{p}, k)$$

  - Minimises suffix errors (maximum-a-posteriori suffix)

- Optimal decision rule for ITP:

$$\hat{\boldsymbol{s}} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N) \qquad \text{for some } N$$

$$\hat{s}_i = \operatorname*{argmax}_{s_i} p(s_i \mid \boldsymbol{x}, \boldsymbol{p}, k, \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{i-1})$$

  - Minimises interactions (maximum-a-posteriori word)
  - Suffix is generated by appending one word at a time

- We knew how to compute:

$$\hat{s}_i = \operatorname*{argmax}_{s_i} \sum_{s'} p(s_i, \boldsymbol{s}' \mid \boldsymbol{x}, \boldsymbol{p}, k, \hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{i-1})$$

  - Efficient (forward-backward) algorithm using wordgraphs

Optimum Decision Rule: Greedy search algorithm

- Transform word graphs into deterministic probabilistic finite state automata
- Efficient algorithm for transformation (opt min.)
- Word graph transformation techniques:

algorithm	no. states (%)		no. arcs (%)		time (ms)	
weighted min.	1146	[20, 2982]	4498	[50, 12477]	27900	[10, 53190]
opt min.	16	[5, 23]	38	[21, 54]	198	[10, 560]

- If $p(s_i|\cdot)$ is sorted then predicting $\boldsymbol{s}$ is $O(N)$

Prediction as a Machine Learning Problem

- Predicting the next word is seen as a classification problem

$$p(s_1 \mid \boldsymbol{x}, \boldsymbol{p}) = \frac{1}{Z} \sum_j \text{feature}_j(s_1, \boldsymbol{x}, \boldsymbol{p}) \tag{1}$$

- Support Vector Machines as classifier

- Baseline: classical ITP with error recovery:
  - edit distance as primary objective
  - path score as second objective

Machine Learning: Setup

- Data
  - post-edited sentences from first field trial (1144 sentences)
  - machine translation system used for first field trial

  → also generate search graphs

- Simulated Task
  - use user translation for simulated user actions
  - task: predict each word in user translation
  - evaluation: ratio (%) of correctly predicted words

- Generate and test
  - consider all partial paths in search graph
  - score all partial paths with model (match to prefix? good path?)
  - select best-scored partial path, predict next word

Machine Learning: Features

- path score of matched prefix (user input) and prediction
- number of states
- average path score (score/states)
- number of deletions (del), the number of insertions (ins) and the number of substitutions (sub) needed to match the prefix
- total edits sed = del+ins+sub
- count averaged by number of tokens of the matched prefix (AvgSed)
- whether the last 1, 2, or 3 tokens were matched (lastMatched, last2Matched, last3Matched)
- levensthein (leven) distance between the last token of the prefix and the matched string (in case it is the same word but in e.g. plural form)
- prefix size
- whether the user input was larger than the matched string

Prediction as a Machine Learning Problem

- Oracle performance
  - correct word in search graph: 79.6%
  - correct word in 100-best list: 71.1%

- Classifier accuracy on 100-best list with possible correct word:

	correct	**false positives**	**false negatives**
All features	1296 (80.30%)	156 (9.66%)	162 (10.04%)
Sed + LastMatched	1341 (83.09%)	135 (8.36%)	138 (8.55%)
Sed + LastMatched + leven	1340 (83.02%)	135 (8.36%)	139 (8.62%)
Sed + LastMatched + sub	1342 (83.15%)	134 (8.30%)	138 (8.55%)
Sed + LastMatched + sub+ins	1345 (83.33%)	161 (9.98%)	108 (6.69%)
Sed + LastMatched + sub+ins+del	1344 (83.27%)	162 (10.04%)	108 (6.69%)

Task 2.2

Multi-Modality in Translation Prediction

Task 2.2: Multi-Modality in Translation Prediction

- Schedule: months 6–24
- Goal: introduce new technologies into the interaction framework
- Interaction with mouse and keyboard is intuitive, but often simple and inefficient
- Introduce other intuitive HCI devices: e-pen or touch-screens
- Focus: design a comfortable and efficient user interface

Context and motivation

- More comfortable and effective user interfaces
- *E-pen* based interaction is a promising alternative to keyboard and mouse
- Recognition errors could reduce productivity w.r.t. keyboard:
  - partial sentences vs full sentences
  - very short contexts: $1 \sim 3$ words per feedback
  - feedback for MT errors $\approx$ low LM probability
- A rule of thumb to user acceptability:
  $< 1\%$ excellent
  $< 3\%$ acceptable
  $5\% \sim 20\%$ acceptable if there is a substantial payoff in terms of achieving task goals.
- Aiming at a more comfortable system:
  - Recognition of sub-word units and sequences of multiple words
  - Recover from HTR errors
  - Recognition of e-pen gestures

On-line HTR results

Sub-word editing

Word and phrase editing

HTR : baseline

4**PREF** : 4-gram prefix

M1/M2 : word model $1/2$

System (en)	CN	CN$_p$	W-CN$_p$	M-CN$_p$
CER (%)	11.4	**3.2**	8.5	8.4

System	es (CER %)		en (CER %)	
	word	phrase	word	phrase
HTR	11.1	16.8	9.9	18.6
n**PREF**	9.9	16.3	9.5	18.0
M1/M2	**8.6**	17.0	7.7	17.4
M1/M2+4**PREF**	9.0	**15.2**	**7.5**	**15.5**

- phrase HTR more complex:
  – segmentation errors, $\approx 30\%$ errors are deletions

- Using n-best lists for word HTR
  – 10-best list: 3.7% (es) and 2.8% (en) error rates

Proof-reading gestures

deletion	$\text{if}_1\ \text{any}_2\ \text{feature}_3\ \ \text{not}_4\ \text{is}_5\ \text{available}_6\ \text{on}_7\ \text{your}_8\ \text{network}_9$
insertion	$\text{if}_1\ \text{any}_2\ \text{feature}_3\ \text{not}_4\ \text{is}_5\ \text{available}_6\ \text{on}_7\ \text{your}_8\ \text{network}_9$
substitution	$\text{if}_1\ \text{any}_2\ \text{feature}_3\ \text{not}_4\ \text{is}_5\ \text{available}_6\ \text{on}_7\ \text{your}_8\ \text{network}_9$
shift	$\text{if}_1\ \text{any}_2\ \text{feature}_3\ \text{not}_4\ \text{is}_5\ \text{available}_6\ \text{on}_7\ \text{your}_8\ \text{network}_9$
transposition	$\text{if}_1\ \text{any}_2\ \text{feature}_3\ \text{not}_4\ \text{is}_5\ \text{available}_6\ \text{on}_7\ \text{your}_8\ \text{network}_9$

TER

Error rate with state-of-the-art gesture recognizer $\approx 10\%$

Study of E-pen Gestures for Text Editing: MinGestures

LABEL	ACTION	RESULT
Substitute	Lorem Ipsum 	Lorem Ipsan
Merge	Lorem Ipsum	LoremIpsum
Delete	Lorem Ipsum 	Lorem
Insert	Lorem Ipsum 	Lorem et Ipsum

LABEL	ACTION	RESULT
Split	Lorem	Lor em
Validate	Lorem Ipsum 	Lorem Ipsum
Undo	Lorem	Lorem Ipsum
Redo	Lorem Ipsum 	Lorem

- Simple implementation
- Fast recognition ($< 1ms$)
- Robust recognition ($\approx 1\%$ error rate)
- Intuitive and easy to remember

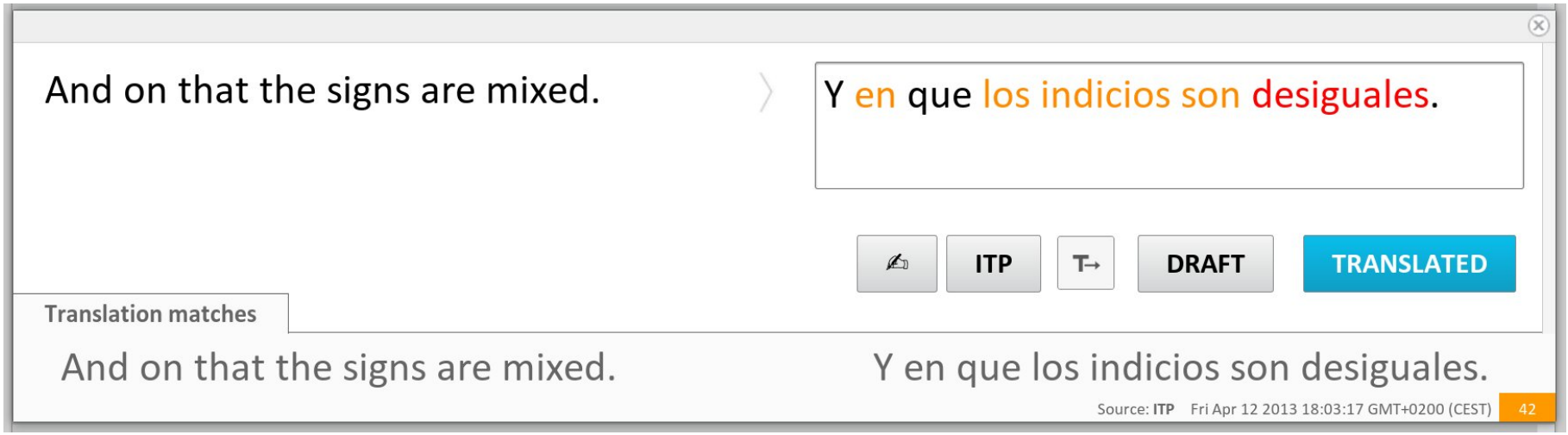
Task 2.3

Prediction from active interaction

Task 3: Prediction from Active Interaction

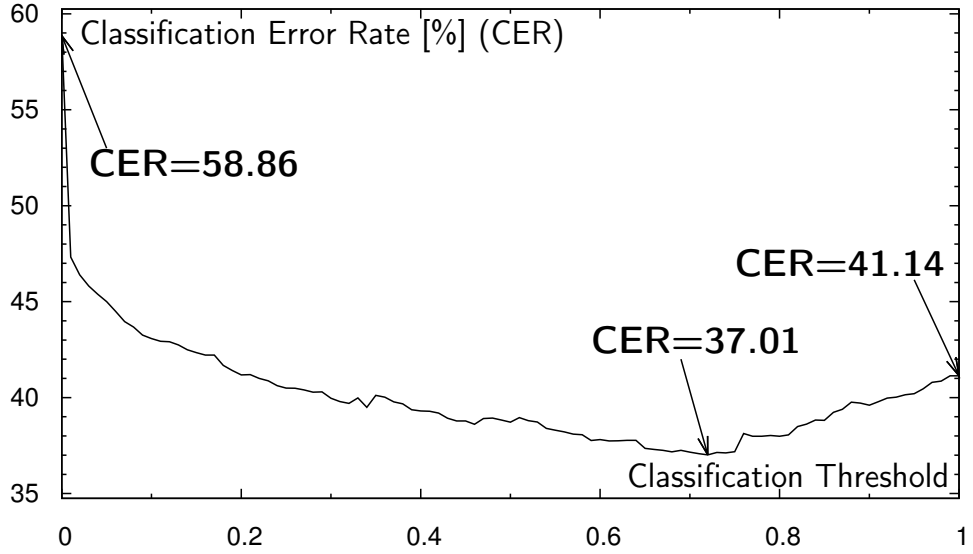
- Schedule: months 13–36
- Goal: aid the user in identifiying incorrect translations
  - The ITP system becomes an active agent of the translation process
- Approach: inform the user about the system's confidence on each translated word
- Focus: study the impact of providing the user with such confidence information
- Active interaction is based on results from WP3.2
- Active interaction influences active learning in WP4.2

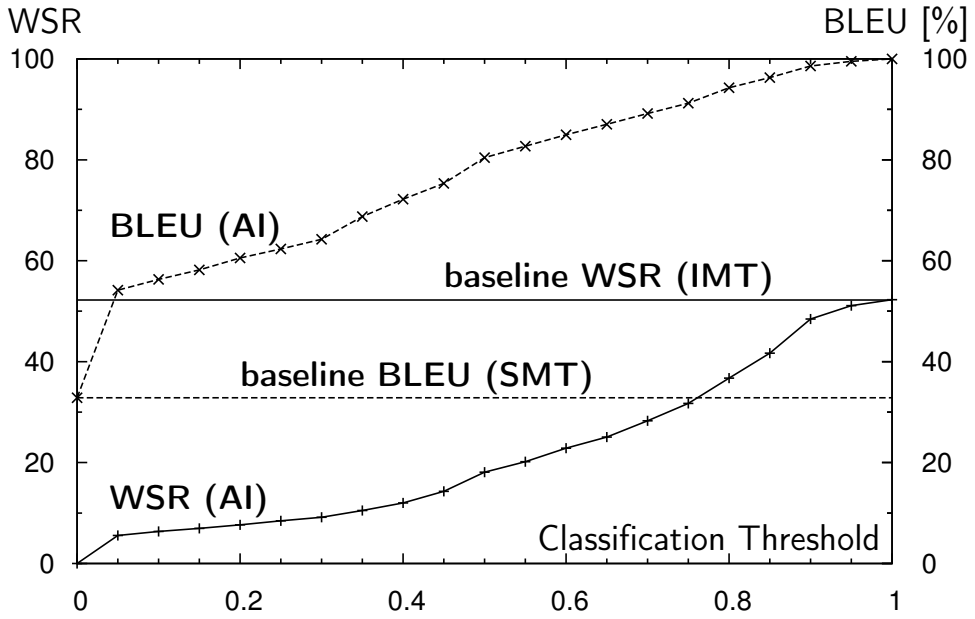
Active Interaction in the CasMaCat Workbench



- Two highlight thresholds:
  - Red: Probable incorrect translations
  - Orange: Dubious translations

In-Laboratory Results





- Confidence scores are more informative that consider all words correct or incorrect
- Active interaction allows for a trade-off between translation quality and human effort

Human Evaluation

- Qualitative experimentation in contrast to the previous quantitative results
  - Users' opinions about the potential of the proposed active interaction protocol
- Users reckon active interaction as a desirable feature but...
  ... they also perceive confidence information as too error-prone:

  "I could definitely benefit from this type of visual aid (active interaction), but the system still needs to make better predictions"
- Different errors have different penalties for the human user
  - In-laboratory experiments penalize equally all errors
- A non uniform penalization is required to match the expectations of the users

Task 2.4

Prediction from Parse Forest

Task 2.4: Prediction from Parse Forest

- Schedule: months 6–18
- Goal: use hierarchical SMT models as a formal approach to manage reordering
- Approaches: we extend conventional ITP technology in two directions
  - Providing a statistical approach to the prefix coverage problem
  - A new formalization to include both phrase-based and hierarchical SMT models
- Focus: Empirical comparison against conventional ITP and decoupled post-editing
  - Human effort required to generate correct translations

Development

- Two new ITP formalizations that include stochastic error correction
  - Error-correction though a statistical interpretation of the edit distance
  - *Conditioned suffix:* the predicted suffix is conditioned by the user-defined prefix
  - *Independent suffix:* the predicted suffix is part of a translation indirectly conditioned by the user defined prefix
- Extend the ITP implementation to use hierarchical SMT models
  - New implementation based on hypergraphs
- Common framework for both phrase-based and hierarchical SMT

Comparison of Different ITP Setups

IMT Setup	EU (Es→En)		TED (Zh→En)	
	phrase-based	hierarchical	phrase-based	hierarchical
Independent suffix form.	27.4±.5	26.5±.5*	53.0±.4	52.3±.4*
Conditioned suffix form.	26.6±.5*	**25.1±.5**★	52.2±.4*	**50.8±.4**★

KSMR [%] scores. Independent suffix using phrase-based model is our baseline

* denotes a better result compared to baseline, ★ denotes better result than all the other setups (99% confidence)

- Conditioned suffix outperformed Independent suffix
- Hierarchical translation model outperformed phrase-based model

Comparison to Post-Edition

EU (Es→En)		TED (Zh→En)	
post-edition	**conditioned suffix hierarchical**	**post-edition**	**conditioned suffix hierarchical**
PKSR [%]	KSR [%]	PKSR [%]	KSR [%]
27.1	**14.1 (48%)**	40.8	**29.7 (27.2%)**

In parenthesis we display the estimated effort reduction of IMT respect to post-edition

- Estimated human effort was significantly reduced when using IMT
- However, effort savings must be taken with caution
  - They may be difficult to achieve by an actual human user

Efficient Algorithms for Approximate Prefix Matching

- Task: Find derivation in parse forest with
  - minimal string edit distance
  - best model score

- Two exact dynamic programming algorithms
  - top-down search, matching user prefix left to right
  - bottom-up search, match constituents in parse forest against prefix

  $\rightarrow$ top-down search is faster

- Refinements
  - reduction of spurious ambiguity (15-30% faster)
  - normalizing non-matched words (not much gain)
  - inside-outside pruning of parse forest (allows speed/quality trade-off)

Conclusions

- Task 2.1:
  - Optimal decision rule reduces time and space requirements in search
  - SVM for word prediction improve accuracy
- Task 2.2: Use e-pen for interacting with PE or ITP system
  - Promising results in character-based, word-base and phrase-based interaction
  - N-best lists make system usable
  - Simple gestures are efficient and accurate but less expressive
- Task 2.3:
  - Confidence measures are liked by users, but
  - State-of-the-art results are not good enough
- Task 2.4: Use parse forest for ITP
  - Improved results over phrase-based ITP
  - Efficient algorithms for prefix matching

Future Work

- Conduct experiments with standard CASMACAT corpora
- Task 2.1: task completed
  - Suffix changes should be limited as hinted by human evaluation (WP1)
  - Continue research on machine learning ITP
- Task 2.2: task completed
  - More expressive, proof-reading gestures
  - Add character-based and phrase-based interaction to CASMACAT
- Task 2.3:
  - Need to improve confidence measures as hinted by human evaluation (WP1)
- Task 2.4: task completed
  - Implement prediction from ITG and SCFG parse forests

Publications

- Daniel Martín-Albo, Verónica Romero, and Enrique Vidal. Interactive off-line handwritten text transcription using on-line handwritten text as feedback. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1312–1316, 2013.
- Vicent Alabau, Alberto Sanchis, and Francisco Casacuberta. Improving on-line handwritten recognition in interactive machine translation. *Pattern Recognition*, 2013. In press.
- Luis A. Leiva, Vicent Alabau, and Enrique Vidal. Error-proof, high-performance, and context-aware gestures for interactive text edition. In *Proceedings of the 2013 annual conference extended abstracts on Human factors in computing systems (CHI EA)*, pages 1227–1232, 2013.
- Jesús González-Rubio, Daniel Ortiz-Martínez, José-Miguel Benedí, and Francisco Casacuberta. Interactive machine translation using hierarchical translation models. In *Proceedings of the conference on Empirical methods on natural language processing*, 2013.