The TALP-UPC Approach to System Selection: ASIYA Features and Pairwise Classification using Random Forests

Lluís Formiga[1], **Meritxell Gonzàlez**[1], **Alberto Barrón-Cedeño**[1,2]
José A. R. Fonollosa[1] and **Lluís Màrquez**[1]
[1] TALP Research Center, Universitat Politècnica de Catalunya, Spain
[2] Facultad de Informática, Universidad Politécnica de Madrid, Spain
{lluis.formiga,jose.fonollosa}@upc.edu, {mgonzalez,albarron,lluism}@lsi.upc.edu

Abstract

This paper describes the TALP-UPC participation in the WMT'13 Shared Task on Quality Estimation (QE). Our participation is reduced to task 1.2 on *System Selection*. We used a broad set of features (86 for German-to-English and 97 for English-to-Spanish) ranging from standard QE features to features based on pseudo-references and semantic similarity. We approached system selection by means of pairwise ranking decisions. For that, we learned Random Forest classifiers especially tailored for the problem. Evaluation at development time showed considerably good results in a cross-validation experiment, with Kendall's τ values around 0.30. The results on the test set dropped significantly, raising different discussions to be taken into account.

1 Introduction

In this paper we discuss the TALP-UPC[1] participation in the WMT'13 Shared Task on Quality Estimation (QE). Our participation is circumscribed to task 1.2, which deals with System Selection. Concretely, we were required to rank up to five alternative translations for the same source sentence produced by multiple MT systems, in the absence of any reference translation.

We used a broad set of features; mainly available through the last version of the ASIYA toolkit for MT evaluation[2] (Giménez and Màrquez, 2010). Concretely, we derived 86 features for the German-to-English subtask and 97 features for English-to-Spanish. These features cover different approaches and include standard Quality Estimation features, as provided by the above mentioned ASIYA toolkit and *Quest* (Specia et al., 2010), but also a variety of features based on *pseudo-references* (Soricut and Echihabi, 2010), explicit semantic analysis (Gabrilovich and Markovitch, 2007) and specialized language models. See section 3 for details.

In order to model the ranking problem associated to the system selection task, we adapted it to a classification task of pairwise decisions. We trained Random Forest classifiers (and compared them to SVM classifiers), expanding the work of Formiga et al. (2013), from which a full ranking can be derived and the best system per sentence identified.

Evaluation at development time, using cross-validation, showed considerably good and stable results for both language pairs, with correlation values around 0.30 (Kendall τ coefficient) classification accuracies around 52% (pairwise classification) and 41% (best translation identification). Unfortunately, the results on the test set were significantly lower. Current research is devoted to explain the behavior of the system at testing time. On the one hand, it seems clear that more research regarding the assignment of ties is needed in order to have a robust model. On the other hand, the release of the gold standard annotations for the test set will facilitate a deeper analysis and understanding of the current results.

The rest of the paper is organized as follows. Section 2 describes the ranking models studied for the system selection problem. Section 3 describes the features used for learning. Section 4 presents the setting for parameter optimization and feature selection and the results obtained. Finally, Section 5 summarizes the lessons learned so far and outlines some lines for further research.

2 Ranking Model

We considered two learning strategies to obtain the best translation ranking model: SVM and Random

[1] Center for Language and Speech Technologies and Applications (TALP), Technical University of Catalonia (UPC).

[2] http://asiya.lsi.upc.edu

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 359–364,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

Forests. Both strategies were based on predicting pairwise quality ranking decisions by means of supervised learning. These decision was motivated from our previous work (Formiga et al., 2013) were we learned that they were more consistent to select the best system (according to human and automatic metrics) compared to absolute regression approaches. In that work we used only the subset of features 1, 2, 3 and 8 described in Section 3. For this shared task we have introduced additional similarity measures (subsets 4 to 7) that feature semantic analysis and automatic alignments between the source and the translations.

The rationale for transforming a ranking problem to a pairwise classification problem has been described previously in several work (Joachims, 2002; Burges et al., 2005). The main idea is to ensemble the features of both individuals and assign a class {-1,1} which tries to predict the pairwise relation among them. For linear based approach this adaptation is as simple to compute the difference between features between all the pairs of the training data.

We used two different learners to perform that task. First, we trained a Support Vector Machine ranker by means of pairwise comparison using the SVM^{light} toolkit (Joachims, 1999), but with the "`-z p`" parameter, which can provide system rankings for all the members of different groups. The learner algorithm was run according to the following parameters: RBF-kernel, expanding the working set by 9 variables at each iteration, for a maximum of 50,000 iterations and with a cache size of 100 for kernel evaluations. The trade-off parameter was empirically set to 0.001. This implementation ignores the ties for the training step as it only focuses in better than/ worse than relations.

Secondly, we used Random Forests (Breiman, 2001), the rationale was the same as ranking-to-pairwise implementation from SVM^{light}. However, SVM^{light} considers two different data preprocessing methods depending on the kernel of the classifier: LINEAR and RBF-Kernel. We used the same data-preprocessing algorithm from SVM^{light} in order to train a Random Forest classifier with ties (three classes: {0,-1,1}) based upon the pairwise relations. We used the Random Forests implementation of scikit-learn toolkit (Pedregosa et al., 2011) with 50 estimators.

Once the classes are given by the Random Forest, we build a graph by means of the adjacency matrix of the pairwise decision. Once the adjacency matrix has been built, we assign the final ranking through a dominance scheme similar to Pighin et al. (2012). In that case, however, there are not topological problems as the pairwise relations are complete across all the edges.

3 Features Sets

We considered a broad set of features: 97 and 86 features for English-to-Spanish (*en-es*) and German-to-English (*de-en*), respectively. We grouped them into the following categories: *baseline QE metrics*, *comparison against pseudo-references*, *source-translation*, and *adapted language models*. We describe them below. Unless noted otherwise, the features apply to both language pairs.

3.1 Baseline Features

The baseline features are composed of well-known quality estimation metrics:

1. *Quest Baseline (QQE)*
   Seventeen baseline features from Specia et al. (2010). This set includes token counts (and their ratio), LM probabilities for source and target sentences, percentage of n-grams in different quartiles of a reference corpus, number of punctuation marks, and fertility ratios. We used these features in the *en-es* partition only.

2. ASIYA*'s QE-based features (AQE)*
   Twenty-six QE features provided by ASIYA (Gonzàlez et al., 2012), comprising bilingual dictionary ambiguity and overlap; ratios concerning chunks, named-entities and PoS; source and candidate LM perplexities and inverse perplexities over lexical forms, chunks and PoS; and out-of-vocabulary word indicators.

3.2 Pseudo-Reference-based Features

Soricut and Echihabi (2010) introduced the concept of pseudo-reference-based features (PR) for translation ranking estimation. The principle is that, in the lack of human-produced references, automatic ones are still good for differentiating good from bad translations. One or more secondary MT systems are required to generate translations starting from the same input, which are taken as pseudo-references. The similarity towards the pseudo-references can be calculated with any evaluation measure or text similarity function, which gives us all feature variants in this group. We consider the following PR-based features:

3. *Derived from* ASIYA*'s metrics (APR)*
   Twenty-three PR features, including GTM-l ($l \in \{1,2,3\}$) to reward different length matching (Melamed et al., 2003), four variants of ROUGE (-L, -S*, -SU* and -W) (Lin and Och, 2004), WER (Nießen et al., 2000), PER (Tillmann et al., 1997), TER, and TER$_{base}$ (i.e., without stemming, synonymy look-up, nor paraphrase support) (Snover et al., 2009), and all the shallow and full parsing measures (i.e., constituency and dependency parsing, PoS, chunking and lemmas) that ASIYA provides either for Spanish or English as target languages.

4. *Lexical similarity (NGM)*
   Cosine and Jaccard coefficient similarity measures for both token and character n-grams considering $n \in [2,5]$ (i.e., sixteen features). Additionally, one Jaccard-based similarity measure for "pseudo-prefixes" (considering only up to four initial characters for every token).

5. *Based on semantic information (SEM)*
   Twelve features calculated with named entity- and semantic role-based evaluation measures (again, provided by ASIYA). Sentences are automatically annotated using SwiRL (Surdeanu and Turmo, 2005) and BIOS (Surdeanu et al., 2005). We used these features in the *de-en* subtask only.

6. *Explicit semantic analysis (ESA)*
   Two versions of explicit semantic analysis (Gabrilovich and Markovitch, 2007), a semantic similarity measure, built on top of Wikipedia (we used the opening paragraphs of $100k$ Wikipedia articles as in 2010).

3.3 Source-Translation Extra Features

Source-translation features include explicit comparisons between the source sentence and its translation. They are meant to measure how *adequate* the translation is, that is, to what extent the translation expresses the same meaning as the source. Note that a considerable amount of the features described in the *baseline* group (*QQE* and *AQE*) fall in this category. In this subsection we include some extra features we devised to capture source–translation dependencies.

7. *Alignment-based features (ALG / ALGPR)*
   One measure calculated over the aligned words between a candidate translation and the source *(ALG)*; and two measures based on the comparison between these alignments for two different translations (e.g., candidate and pseudo-reference) and the source *(ALGPR)*.[3]

8. *Length model (LeM)*
   A measure to estimate the quality likelihood of a candidate sentence by considering the "expected length" of a proper translation from the source. The measure was introduced by (Pouliquen et al., 2003) to identify document translations. We estimated its parameters over standard MT corpora, including Europarl, Newswire, Newscommentary and UN.

3.4 Adapted Language-Model Features

We interpolated different language models comprising the WMT'12 Monolingual corpora (EPPS, News, UN and Gigafrench for English). The interpolation weights were computed as to minimize the perplexity according to the WMT Translation Task test data (2008-2010)[4]. The features are as follow:

9. *Language Model Features (LM)*
   Two log-probabilities of the translation candidate with respect to the above described interpolated language models over word forms and PoS labels.

4 Experiments and Results

In this section we describe the experiments carried out to select the best feature set, learner, and learner configuration. Additionally, we present the final performance within the task. The setup experiments were addressed doing two separate 10-fold cross validations on the training data and averaging the final results. We evaluated the results through three indicators: Kendall's τ with no

[3] Alignments were computed with the Berkeley aligner https://code.google.com/p/berkeleyaligner/

[4] http://www.statmt.org/wmt13/translation-task.html

penalization for the ties, accuracy in determining the pairwise relationship between candidate translations, and global accuracy in selecting the best candidate for each source sentence.

First, we compared our SVM learner against Random Forests with the two variants of data preprocessing (LINEAR and RBF). In terms of Kendall's τ, we found that the Random Forests (RF) were clearly better compared to SVM implementation. Concretely, depending on the final feature set, we found that RF achieved a τ between 0.23 and 0.29 while SVM achieved a τ between 0.23 and 0.25. With respect to the accuracy measures we did not find noticeable differences between methods as their results moved from 49% to 52%. However, considering the accuracy in terms of selecting only the best system there was a difference of two points (42.2% vs. 40.0%) between methods, being RF again the best system. Regarding the pairwise preprocessing the results between RBF and LINEAR based preprocessing were comparable, being RBF slightly better than LINEAR. Hence, we selected Random Forests with RBF pairwise preprocessing as our final learner.

de-en	**τ with ties**		***Accuracy***	
	Ignored	***Penalized***	***All***	***Best***
AQE+LeM+ALGPR+LM	***33.70***	***15.72***	***52.56***	***41.57***
AQE+SEM+LM	***32.49***	***14.61***	***52.72***	***40.92***
AQE+LeM+ALGPR+ESA+LM	32.08	13.81	52.71	41.37
AQE+ALG+ESA+SEM+LM	32.06	13.96	52.20	40.64
AQE+ALG+LM	31.97	14.29	52.00	40.83
AQE+LeM+ALGPR+SEM+LM	31.93	13.57	52.52	40.98
AQE+ESA+SEM+LM	31.79	13.68	52.50	40.76
AQE+LeM+ALGPR+ESA+SEM+LM	31.72	14.01	52.65	40.83
AQE+ALG+SEM+LM	31.17	12.86	52.18	40.51
AQE+ALG+SEM	30.72	12.58	51.75	39.66
AQE+LeM+ALGPR+ESA+SEM	30.47	11.79	51.85	39.58
AQE+ESA+LM	30.31	12.23	52.60	40.69
AQE+ALG+ESA+LM	30.26	12.40	52.03	40.99
AQE+LeM+ALGPR	30.24	11.83	51.96	40.42
AQE+LeM+ALGPR+SEM	30.23	11.84	52.10	40.32
AQE+LeM+ALGPR+ESA	29.89	11.87	51.83	40.07
AQE+ALG+ESA	29.81	11.30	51.37	39.47
AQE+SEM	29.80	12.06	51.75	39.52
AQE+NGM+APR+ESA+SEM+LM	29.34	10.58	51.33	38.55
AQE+ESA+SEM	29.31	11.46	51.66	39.24
AQE+ESA	29.13	11.12	51.82	39.90
AQE+ALG+ESA+SEM	28.35	10.32	51.37	38.98
AQE+NGM+APR+ESA+SEM	27.55	9.22	51.01	38.12

Table 1: Set-up results for *de-en*

For the feature selection process, we considered the most relevant combinations of feature groups. Table 1 shows the set-up results for the *de-en* subtask and Table 2 shows the results for the *en-es* subtask.

In terms of τ we observed similar results between the two language pairs. However accuracies for the *de-en* subtask were one point above the ones for *en-es*. Regarding the features used, we found that the best feature combination to use was composed of: *i*) a baseline QE feature set (Asiya or Quest) but not both of them, *ii*) Length Model, *iii*) Pseudo-reference aligned based features and the use of *iv*) adapted language models. However, within the *de-en* subtask, we found that substituting Length Model and Aligned Pseudo-references by the features based on Semantic Roles (SEM) could bring marginally better accuracy. We also noticed that the learner was sensitive to the features used so selecting the appropriate set of features was crucial to achieve a good performance.

en-es	**τ with ties**		***Accuracy***	
	Ignored	***Penalized***	***All***	***Best***
QQE+LeM+ALGPR+LM	***33.81***	***15.87***	***51.66***	***41.01***
AQE+LeM+ALGPR+LM	***33.75***	***16.44***	***51.56***	***41.52***
QQE+AQE+LM	32.71	14.59	51.18	41.02
QQE+AQE+LM+ESA	32.69	15.30	51.48	41.30
QQE+AQE+LeM+ALGPR+LM+ESA	32.63	13.64	51.39	40.48
QQE+AQE+LeM+ALGPR+LM	32.41	14.06	51.43	40.49
QQE+LeM+ALGPR+LM+ESA	31.66	13.39	51.37	41.05
QQE+AQE+ALG+LM	31.46	13.62	51.28	41.29
AQE+LeM+ALGPR+LM+ESA	31.29	14.10	51.55	41.43
QQE+AQE+ALG+LM+ESA	31.25	13.58	51.64	41.66
QQE+AQE+NGM+APR+LM+ESA	30.58	12.48	50.93	40.66
QQE+AQE+NGM+APR+LM	29.94	12.54	50.95	40.25
QQE+AQE	28.98	10.92	49.97	39.65
QQE+AQE+LeM+ALGPR	28.94	10.48	49.99	39.71
QQE+AQE+NGM+ESA+LM	28.85	11.88	50.90	40.22
AQE+LeM+ALGPR	28.81	10.11	50.06	40.01
QQE+AQE+ESA	28.68	10.31	49.96	39.27
AQE+ESA	28.67	10.81	50.35	39.18
AQE	28.65	10.68	49.76	38.90
QQE+AQE+ALG	28.47	9.63	49.67	39.66
QQE+AQE+NGM+APR+ESA	28.43	9.75	49.67	38.74
QQE+AQE+NGM	27.23	9.10	49.44	38.98
QQE+AQE+ALG+ESA	27.08	7.93	50.26	39.71
QQE+AQE+LeM+ALGPR+ESA	27.03	8.65	50.35	40.49
AQE+LeM+ALGPR+ESA	26.96	8.26	50.30	39.47
QQE+AQE+NGM+ESA	26.59	7.56	49.52	38.62
QQE+AQE+NGM+APR	25.39	6.97	49.90	39.53

Table 2: Setup results for *en-es*

de-en	**τ (ties penalized,**
ID	**non-symmetric between [-1,1])**
Best	0.31
UPC AQE+SEM+LM	0.11
UPC AQE+LeM+ALGPR+LM	0.10
Baseline Random-ranks-with-ties	-0.12
Worst	-0.49

Table 3: Official results for the *de-en* subtask (ties penalized)

en-es	**τ (ties penalized,**
ID	**non-symmetric between [-1,1])**
Best	0.15
UPC QQE+LeM+ALGPR+LM	-0.03
UPC AQE+LeM+ALGPR+LM	-0.06
Baseline Random-ranks-with-ties	-0.23
Worst	-0.63

Table 4: Official results for the *en-es* subtask (ties penalized)

In Tables 3, 4, 5 and 6 we present the official results for the WMT'13 Quality Estimation Task, in all evaluation variants. In each table we compare to the best/worst performing systems and also to the official baseline.

We can observe that in general the results on the test sets drop significantly, compared to our

de-en	τ (ties ignored,	Non-ties
ID	symmetric between [-1,1])	/ (882 dec.)
Best	0.31	882
UPC AQE+SEM+LM	0.27	768
UPC AQE+LeM+ALGPR+LM	0.24	788
Baseline Random-ranks-with-ties	0.08	718
Worst	-0.03	558

Table 5: Official results for the *de-en* subtask (ties ignored)

en-es	τ (ties ignored,	Non-ties
ID	symmetric between [-1,1])	/ (882 dec.)
Best	0.23	192
UPC QQE+LeM+ALGPR+LM	0.11	554
UPC AQE+LeM+ALGPR+LM	0.08	554
Baseline Random-ranks-with-ties	0.03	507
Worst	-0.11	633

Table 6: Official results for the *en-es* subtask (ties ignored)

set-up experiments. Restricting to the evaluation setting in which ties are not penalized (i.e., corresponding to our setting during system and parameter tuning), we can see that the results corresponding to *de-en* (Table 5) are comparable to our set-up results and close to the best performing system. However, in the *en-es* language pair the final results are comparatively much lower (Table 6). We find this behavior strange. In this respect, we analyzed the inter-annotator agreement within the gold standard. Concretely we computed the Cohen's κ for all overlapping annotations concerning at least 4 systems for both language pairs. The results of our analysis are presented in Table 7 and therefore it confirms our hypothesis that en-es annotations had more noise providing an explanation for the accuracy decrease of our QE models and setting the subtask into a more challenging scenario. However, further research will be needed to analyze other factors such as oracles and improvement on automatic metrics prediction and reliability compared to linguistic expert annotators.

Another remaining issue for our research concerns investigating better ways to deal with ties, as their penalization lowered our results dramatically. In this direction we plan to work further on the adjacency matrix reconstruction heuristics and presenting the features to the learner in a structured form.

# of systems	Lang	Cohen's κ	# of elements
4	en-es	0.210	560
	de-en	0.369	640
5	en-es	0.211	130
	de-en	0.375	145

Table 7: Golden standard test set agreement coefficients measured by Cohen's κ

5 Conclusions

This paper described the TALP-UPC participation in the WMT'13 Shared Task. We approached the Quality Estimation task based on system selection, where different systems have to be ranked according to their quality. We derive a full ranking and identify the best system per sentence on the basis of Random Forest classifiers.

After the model set-up, we observed considerably good and robust results for both translation directions, German-to-English and English-to-Spanish: Kendall's τ around 0.30 as well as accuracies around 52% on pairwise classification and 41% on best translation identification. However, the results over the official test set were significantly lower. We have found that the low inter-annotator agreement between users on that set might provide an explanation to the poor performance of our QE models.

Our current efforts are centered on explaining the behavior of our QE models when facing the official test sets. We are following two directions: *i*) studying the ties' impact to come out with a more robust model and *ii*) revise the English-to-Spanish gold standard annotations in terms of correlation with automatic metrics to facilitate a deeper understanding of the results.

Acknowledgments

Acknowledgements

This work has been partially funded by the Spanish *Ministerio de Economía y Competitividad*, under contracts TEC2012-38939-C03-02 and TIN2009-14675-C03, as well as from the European Regional Development Fund (ERDF/FEDER) and the European Community's FP7 (2007-2013) program under the following grants: 247762 (FAUST, FP7-ICT-2009-4-247762) and 246016 (ERCIM "Alain Bensoussan" Fellowship).

References

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender.

2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM.

Lluís Formiga, Lluís Màrquez, and Jaume Pujantell. 2013. Real-life translation quality estimation for mt system selection. In *Proceedings of 14th Machine Translation Summit (MT Summit)*, Nice, France, September. EAMT.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Meritxell Gonzàlez, Jesús Giménez, and Lluís Màrquez. 2012. A graphical interface for mt evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144, Jeju Island, Korea, July. Association for Computational Linguistics.

Thorsten Joachims, 1999. *Advances in Kernel Methods – Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT Press.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In ACM, editor, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain, July.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *HLT-NAACL*.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for mt research. In *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC 2000)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

Daniele Pighin, Lluís Formiga, and Lluís Màrquez. 2012. A graph-based strategy to streamline translation quality assessments. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA'2012)*, San Diego, USA, October. AMTA.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.

Matthew G. Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2):117–127.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine Translation Evaluation Versus Quality Estimation. *Machine Translation*, 24:39–50, March.

Mihai Surdeanu and Jordi Turmo. 2005. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of CoNLL Shared Task*.

Mihai Surdeanu, Jordi Turmo, and Eli Comelles. 2005. Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceedings of European Conference on Speech Communication and Technology*.