Tree-Structured Named Entities Extraction from Competing Speech Transcriptions

Davy Weissenbacher[1] and Christian Raymond[1]

(1) INSA - IRISA, INRIA de Rennes
20 Avenue des buttes de Coësmes, Rennes, France
davy.weissenbacher@inria.com, christian.raymond@irisa.fr

Abstract. When real applications are working with automatic speech transcription, the first source of error does not originate from the incoherence in the analysis of the application but from the noise in the automatic transcriptions. This study presents a simple but effective method to generate a new transcription of better quality by combining utterances from competing transcriptions. We have extended a structured Named Entity (NE) recognizer submitted during the ETAPE Challenge. Working on French TV and Radio programs, our system revises the transcriptions provided by making use of the NEs it has detected. Our results suggest that combining the transcribed utterances which optimize the F-measures, rather than minimizing the WER scores, allows the generation of a better transcription for NE extraction. The results show a small but significant improvement of 0.9% SER against the baseline system on the ROVER transcription. These are the best performances reported to date on this corpus.

Index Terms: speech transcription, structured named entities, multi-pass decoding.

When real applications are working with automatic speech transcription, the first error does not originate from the incoherence in the analysis of the application, but from the noise of the automatic transcription outputs. With a rate often close to one in three words incorrect in the transcription, the quality of the preprocessing is low and, as a result, the output analysis of the application is often unexploitable. An explanation for this low performance of speech recognizers can be found in [8]. Little lexical and syntactic information is effectively used to enable the computation of the decoding of the acoustic output. More complex information are reintegrated in a second decoding pass where only the best sequences of words produced during the first pass are considered.

The main contribution of this study is to present a simple but effective method to generate a new transcription of better quality by combining several competing transcriptions. Current Automatic Speech Recognition (ASR) systems rely on various strategies and/or resources to discover the original utterances pronounced. As a consequence, errors made by competing ASRs are different, which make the transcriptions complementary. The Rover method exploits such complementarity to recombine several transcriptions and output a

new transcription [3]. Previous studies to recombine the transcriptions focus on minimizing the Word Error Rate (WER) measure[1].We claim that the WER measure is not the measure of importance and should be ignored [4]. The measure that is more important is the measure of the performance of the system on the final application, and that is the one to be optimized.

To test this hypothesis we have run experiments on structured Named Entity (NE) extraction using the corpus released during the recent ETAPE Challenge. This challenge aimed to evaluate the state of the art in NE extraction on automatic speech transcription of TV and radio French programs. We found promising results, with the best performances achieved to date on this corpus.

In section 1, we first describe the task and the corpus of the evaluation campaign ETAPE, and then provide an overview of the system submitted to extract structured NEs and which ranked first during the campaign. Section 2 details our first investigation to use the NEs extracted to recombine the complementary transcriptions. We report the gain observed during our experiments and the perspectives of this work in section 3.

1 The ETAPE Challenge

The goal of the ETAPE challenge in 2012 was to extract named entities (NEs) from automatic transcription output[2]. The ETAPE corpus [5] consists of 13.50 hours of radio news broadcast and 28.40 hours of TV shows. The corpus was chosen to be difficult to process, with the programs in French language chosen not only from French channels, but also from Moroccan and African radio stations. The programs were selected to include mostly non planned speech and reasonable proportions of conversations with multiple speakers. The data was split into 8.20, 25.50 and 8.20 hours for development, training and testing respectively. Five speech recognizers have been applied on the corpus. Their performances on our test data range from 23% to 35% WER.

The originality of the ETAPE challenge was in its definition of the NEs [18]. A NE is a rigid designator [9], like a proper name or a company name, and is commonly viewed as a simple object, that is a sequence of words. However, a NE can also be seen as a structured object. According to the definition of the ETAPE challenge, NEs have a tree structure and are both hierarchical and compositional. For instance, type *pers* (person) is split into two subtypes, *Pers.ind* (individual person) and *Pers.coll* (collective person). *Pers* entities are composed like in the individual person *Nicolas Sarkozy* where *Nicolas* is the first name and *Sarkozy* the last name. Figure 1 enumerates the 7 main types and the 32 subtypes of the taxonomy. Figure 2 shows all the components.

Learning trees from data is known to be a difficult task. Since complexity issues rise quickly, learning the full tree in one step is often impossible [13].

[1] $WER = \frac{S+D+I}{N}$, where D, I, S stand for the number of deletions, insertions, substitutions of words and N for the total number of words in the reference.

[2] More information about the challenge can be found at *www.afcp-parole.org/etape/workshop.html*

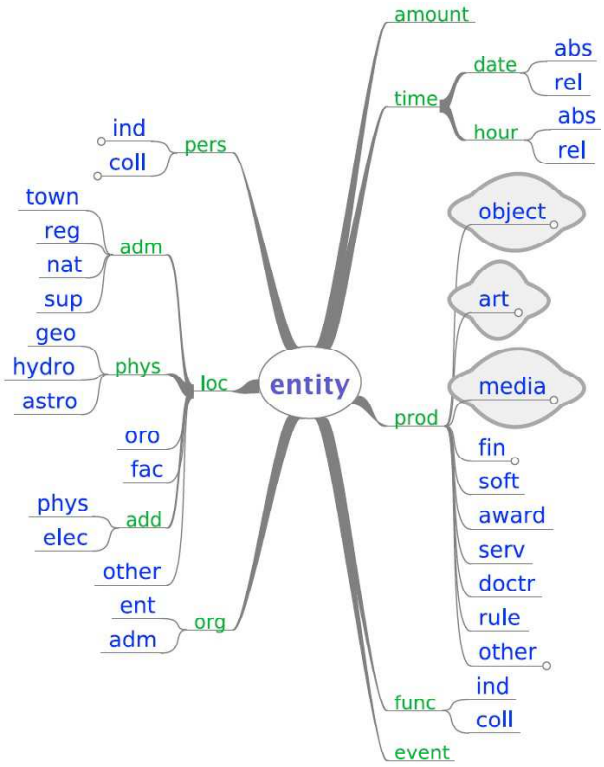

Fig. 1. Named entity hierarchy

Standard approaches to build trees, as grammar and formal based approaches, fail to operate on noisy inputs like automatic speech transcriptions. In contrast, statistical approaches have been proven to be very efficient on both clean and noisy texts. To the best of our knowledge, most statistical NE recognition systems deal with the structure thanks to cascade approaches [15, 1]. But the cascade methodology has an important limitation, errors made in the early stages are propagated through the whole process. The propagation of errors particularly problematic when the inputs are very noisy.

The winning system of the ETAPE challenge avoids the cascade approach by building the trees in two steps. In the first step, it extracts the nodes of all possible trees which may be contained in an utterance. The detection of all the nodes is performed independently in order to avoid complexity issues. For detecting the nodes Conditional Random Fields (CRFs) are used. This sequence labeler is currently one of the best statistical frameworks for NE recognition [12]. Each CRF is trained to recognize a unique type of node, resulting in a total of 68 binary CRFs. Segmentation and labelling are performed in the same time using the *BIO* annotation format. Each CRF uses a common set of features: the words themselves, their associated Part-Of-Speech tags and the mentions

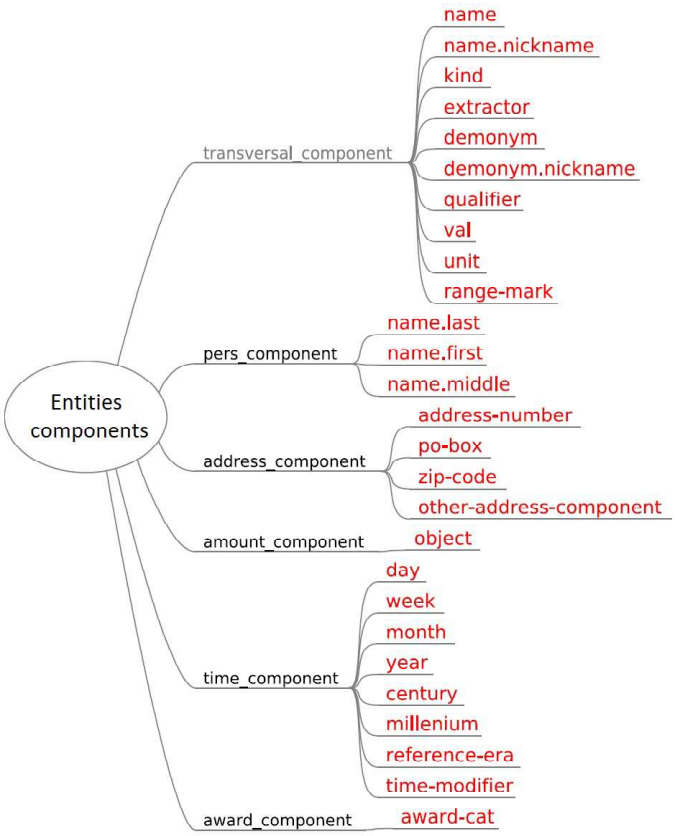

Fig. 2. Entities components

of predefined classes like cities, countries etc. These mentions are extracted by utilizing dictionaries [17].

In the second step the trees are rebuilt from the nodes extracted in the previous step. The simplest and most effective way to rebuild the tree is to choose the nodes of the best analysis for all binary CRFs. To reconstruct a coherent tree one needs to know the subsumption relations between nodes, such as nodes *Pers.ind* always dominating nodes *First.Name*. These relations are learned from the training data. Since nodes are extracted independently by the CRFs, incoherences between their segmentations may occurred. Simple heuristics are employed to recover coherence between erroneous nodes annotations. Despite of its simplicity this algorithm distinctly ranked first among eight participants during the ETAPE challenge with a score of 55.51% WER on the ROVER transcription [16].

Having presented the NE recognizer, in the next section we turn to its use for revising existing transcriptions with the goal of improving the NE extraction scores on the ETAPE corpus.

2 Automatic Transcription Revision Driven by NE Recognition

In this study, we claim that when revising existing transcriptions, the measure of the final task should be optimized rather than minimizing the WER of the new transcription. To improve the quality of the transcription, we select the transcription which maximizes the F-measure of our NE recognizer from the competing transcriptions of an utterance. Considering the same utterance transcribed by two ASRs, the underlying idea is that if a structured NE is recognized in the first transcription and not in the second, it is more likely that the first transcription is correct. Since the final application makes only use of the NEs, the overall quality of the transcription in terms of WER doesn't have to be perfect as long as the NEs can be discovered by the NE recognizer.

We now explain the algorithm followed to generate a new transcription from transcriptions of competing ASRs. Our algorithm takes as input a set of transcriptions output by several ASRs. We have segmented the utterances of all transcriptions to avoid complexity problems[3]. Considering an utterance U_i in the gold transcription, we call the transcriptions output by all competing ASRs for this utterance the *set of competing utterances for* U_i. Each competing utterance in a given set are aligned with the longest utterance in the set using the SCLite algorithm[4]. Each set of competing utterances is then processed sequentially to find the best utterance for each set of competing utterances. In order to select the best utterances for a given set, we apply the NER approach described in the previous section on all competing utterances of the set. The annotated competing utterances are then passed to a Machine Learning (ML) system. The ML, described in the section below, was trained to recognize the best utterance based on the presence or absence of NEs in the utterances. When all best utterances are selected, they are merged to generate a new transcription of the document. The quality of this new transcription is finally evaluated using the official tools provided during the ETAPE challenge for evaluating the original competing transcriptions.

Our algorithm can be illustrated on two competing transcriptions of the following utterances: *U1*, *U1'* for the first set and *U2*, *U2'* for the second set:

***Reference*:** *nous sommes ensemble pour soixante minutes une heure au coeur de tout ce qui fait l'actualité*
[we are together for sixty minutes one hour at the heart of everything which make the news]

[3] As a first working hypothesis, we have segmented the transcriptions based on the gold standard utterances.
[4] SCLite: *www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm*

U1: *nous sommes ensemble pour soixante minutes une heure au coeur de l'actualité*
[we are together for sixty minutes one hour at the heart of the news]
U1': *nous sommes ensemble pour soixante minutes une trop grande tout ce qui fait l'actualité*
[we are together for sixty minutes a too big everything which make the news]
***Reference*:** *c'est ce qu'a dit le ministre Bruno Le Maire ministre de l'agriculture*
[this is what the minister Bruno Le Maire minister of the agriculture said]
U2: *c'est ce qu'a dit la ministre de l'agriculture*
[this is what the minister of the agriculture said]
U2': *c'est ce qu'a dit le ministre Bruno Lemaire ministre de l'agriculture*
[this is what the minister Bruno Lemaire minister of the agriculture said]

The algorithm has to select between *U1* and *U1'* in the first set, and between *U2* and *U2'* in the second. Our algorithm retains utterances with the maximum number of correct NEs. In the first set it is straightforward to select *U1*. It is possible to extract a NE from *U1*, *<Amount> <Val> une </Val> <Unit> heure </Unit> </Amount>*, but not from *U1'* by applying our NE Recognizer on each utterance. Therefore, *U1* is selected as best utterance for the first set. The choice for the second set is less obvious since *U2* and *U2'* both contain NEs. The algorithm has to arbitrate based on the quality of each NE. The problem has been formulated as a classification problem to optimize the decision. An ideal ML framework should prefer *U2'* against *U2*, as the NE *<Func.ind> ministre <Pers.ind> <First.name> Bruno </First.name> <Last.name> Lemaire </Last.name> </Pers.ind> </Func.ind>* is longer and perfectly valid. Once all utterances have been selected, a new transcription composed of *U1* and *U2'* is output. This new transcription is ready for being evaluated on the structured NE extraction task. Since the presence of the NEs is optimized in all utterances of the new transcription, rather than the WER, better performance is expected for the NER task when using this new transcription.

The selection of the best utterance is a ML problem which can be expressed in different ways. We describe here the ML frameworks studied in this work.

Transcriptions Classification. In this framework all competing utterances are submitted to a multi-class classifier. Features available allow the classifier to describe and compare the utterances in order to choose one among them. The features employed here are explained in the table 1. The ML framework which gave the best results on the training corpus was a Bayesian Network. The structure and the conditional probabilities of the Bayesian Network were learned automatically.

Transcriptions Regression. Another framework is to learn directly the F-measure of an utterance using a regression classifier. The selection of the best utterance is done afterwards by picking up the utterance exhibiting the highest F-measure estimation. A bagging-Regression tree for regression obtained the highest performances on our training data and has been chosen for our test. Features used for the regression were similar to those for the *Transcription Classifier*.

Phrases Classification. The features of the previous ML systems provide a global description of the utterances, but this level may appear to be too broad

for our purpose. Not only does the system have to find the utterance containing the maximum number of NEs, but it also has to ensure that the NEs obtained have appropriate qualities. For that reason, we redesigned our system to be able to describe and evaluate independently all phrases annotated as NE in an utterance. The utterance containing the highest number of selected NEs is kept as best utterance.

Let us consider our previous example. When a phrase is annotated as NE by at least one CRF, the phrase and all corresponding phrases in the competing utterances become subject to decision. In *U1*, the phrases *soixante minutes* and *une heure* have been found to be NEs. According to the SCLite alignment, the corresponding phrases in *U1'* are *soixante minutes* and *une trop*. Only *soixante minutes* has been annotated as possible NE in *U1'*. The annotation nodes of *soixante minutes* are the same for *U1* and *U1'*. The choice only relies on the decision taken by the algorithm for the quality of the annotations of *une heure* and *une trop*. The algorithm based on a ML inference gives better credit to the phrase *une heure*, which qualifies the transcription *U1* with two NEs selected against one in *U1'*.

The main component in our algorithm is the ML model used to gauge the phrases. We opt for a multi-class classification to select the best phrases among the competing phrases of each set of utterances. We did not change the ML framework and continue to train a Bayesian Network in the same way as in the previous experiences. The features of table 1 were adapted for describing phrases and completed with the features of the table 2.

Oracle and Baseline Systems. To reveal the maximum improvement possible with our approach, we have computed the performance of an oracle. For all NEs discovered in an utterance by our NER, the oracle is informed with the true value of the NEs. Therefore, it always outputs the best possible utterances for each set of competing utterances given the NE resolution. As a baseline system we have chosen the ROVER transcription. This baseline is a strong baseline since the winning NER achieved its best performances on the ROVER transcription during the challenge.

3 Results and Discussion

In Table 3, we report the NE recognition scores of our system for each recombined transcription given by the ML framework tested. Standard measures of Precision and Recall are completed by the Slot Error Rate (SER) [10], a measure similar to WER which also considers errors made for the segmentation and the labelling of NEs. Both *Transcriptions* and *Phrases* classifiers output a transcription of better quality than the baseline ROVER transcription for our final task. The improvement of 0.9% is shown to be statistically significant with a one-tailed t-test with a degree of liberty = 28 and $\alpha = 0.1$.

These results demonstrate the interest of maximizing the F-measure over minimizing the WER measure when recombining competing transcriptions. The improvement of the recombined transcriptions in terms of WER is not impor-

Table 1. Features describing the set of the 6 Transcriptions in competition.

Feature	Description
CRFs score (for transcription i)	A global score computed by summing all binaries CRFs' probabilities of the words
Max Nodes	The name of the transcription containing the highest number of nodes
Impossible bigram (for transcription i)	The number of sequence of two words never co-occurring in the training corpus
Length (for transcription i)	Total number of words in the transcriptions
Min/Max CRFs scores	The CRFs score of the node which is found to be the min/max score in the competing transcriptions
Mean Node scores	The mean of CRFs scores in the competing transcriptions

tant. The ROVER transcription exhibits a WER of 39.0% whereas the recombined transcriptions produced by the *Transcription classification* shows a WER of 39.2%. This finding corroborates the findings of the ETAPE challenge. NE Recognition performances are not necessarily better on transcriptions with lower WER. When comparing the scores of our system on two transcriptions of the ETAPE data, we found a score of 63.4% SER on the first transcription with 24% WER, and a score of 62.53% SER when the second transcription's WER is of 25% [16]. That is, the SER diminishes by 0.83% whereas the WER increases of 1%.

Analysis of the *Transcription classifier* model informs that this classifier tends to select the ROVER classification of an utterance by default, except when another transcription of the utterance is found with a higher CRF score going along with a higher number of nodes in the utterance. This confirms our intuition: the detection of the NEs is possible only when the transcription reaches a certain threshold of quality and this, in turn, reveals the best transcription among the candidate transcriptions.

A surprising result is the counter-performance of the *Regression* system. This system takes more risks by often picking up utterances that are different from the most reliable ones (*i.e.* ROVER or s23). Although rewarded by a higher recall, it is punished by a drop of precision. The opposite phenomenon is noticed for the *Phrase* classification system. The description of phrases allows the system to discriminate the expected ones and increase its precision with a slight drop of its recall.

Table 2. Features describing phrases in competing transcriptions.

Feature	Description
Max Depth	The size of the longest branch of the structured NE covering the phrase
Max score node	The node which has the highest CRFs score
Existing phrase	1 if an occurrence of the phrase has been found in the training data
Phrase coherence	1 if the phrase is covered by a node known to be a subtype in Fig. 1

Table 3. Performances of NE recognition on recombined transcriptions, in term of Slot Error Rate.

	SER	Precision	Recall
Baseline Rover	.563	.734	.449
Transcriptions Classification	**.554**	.728	.461
Transcriptions Regression	.640	.586	**.463**
Phrases Classification	**.554**	**.738**	.454
Oracle	.509	.751	.499

4 Related Work

A significant number of errors of ASR systems are caused by the Out-Of-Vocabulary words (OOV) since ASR systems rely on a finite lexicon to interpret phonetic inputs[19]. Due to the nature of most of the NEs, that is the open class of Proper Nouns, a large proportion of OOV are unknown NEs. Therefore, a considerable amount of literature has been published on OOV-NEs detection and revision. To date, two complementary approaches have been explored.

The method proposed in this paper is close to the first approach which extends the search space by exploiting multiple sources of information. The simple method is to use multiple ASR system transcriptions as in [3], which results in an important improvement of the WER. More sophisticated methods, with a cost of higher computation complexity, introduce NEs hypotheses directly in the decoding model. In their seminal article, [4] encode the output of a NE recognizer into the loss function of a Minimum Bayes-Risk Classifier to reorder a N-Best list of transcriptions. In a study which worked on a corpus similar to our own [6], [2] make use of the release time of the news to enrich the list of NEs available to the system by adding an external list of NEs known to occur in the documents published during this period of time.

The second type of approach targets specifically the strange grammatical constructions caused by the presence of OOV words with the aim of identifying the

underlying NE(s) [14]. At the last resort, when no NE can be found, a phonetic transcription of the OOV is generally suggested. In 2006, [7] investigated the interest of training a classifier to recognize distortions caused by the unknown NEs. More recently, by noticing that not only do the OOVs deteriorate the transcription at their position in the utterance, but also the immediate context where they appear, [11] rebuilt the best parse of the utterance from a word confusion network with the distinct mentions of OOVs. These latter methods are not in contradiction to our approach, but complement it. While we have empirically established that if the expected NEs occurred in one competing translations, our algorithm will more likely find it, its strong limitation lies in the cases when NEs are absent in the translations. In such cases, nothing can be done to recover and the system relies on the ROVER transcriptions. A module implementing the latter algorithm may be able to detect a NE position and pass to our own system an anonymous NE (or an attempted assertion of the unknown NE) in order to help our system to output the best transcription.

5 Conclusion

Our findings emphasize the interest of optimizing the measure of the final task to improve the quality of the transcription when complementary transcriptions are available. Building on our achievements during the ETAPE Challenge, we have used the structured NEs detected by our NE recognizer to drive the revision of the transcriptions provided to the system. Our results suggest that selecting the competing transcription of the utterances by optimizing the F-measure leads to a better global transcription for NE extraction compared to selection based on a lower WER.

Taking into account the difficulty of the corpus, the results obtained are mainly positive, with a small but significant improvement of 0.9% SER on the recombined transcription against the ROVER baseline. There is, however, still a lot of room for improvement. A promising approach is to recombine the transcriptions by merging all the best transcriptions of phrases, and this even if two distinct phrases of the same utterance belong to different transcriptions. This method will be somehow similar to Word Confusion Network based methods which already have been demonstrated to provide better recombined transcriptions compared to a simple N-Best list recombination [20]. To obviate the unrecoverable limitation when expected NEs do not occur in any transcriptions, as further work, we are considering to integrate a procedure to detect unreliable sequences of words caused by OOV-NEs. This will enable a dedicated algorithm to track down the hidden NEs within external resources before the recombining stage of the transcriptions.

Acknowledgments

We thank Dr. Abeed Sarker and Dr. Graciela Gonzalez for their helpful comments and remarks.

References

1. Dinarelli, M., Rosset, S.: Models cascade for tree-structured named entity detection. In: Proceedings of International Joint Conference on Natural Language Processing (IJCNLP). pp. 1269–1278 (2011)
2. Favre, B., Béchet, F., Nocéra, P.: Robust named entity extraction from large spoken archives. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). pp. 491–498 (2005)
3. Fiscus, J.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In: Proceedings IEEE Automatic Speech Recognition and Understanding Workshop. pp. 347–352 (1997)
4. Goel, V., Byrne, W.: Minimum bayes-risk automatic speech recognition. Computer Speech and Language 14(2), 115–135 (2000)
5. Gravier, G., Adda, G., Paulson, N.and Carré, M., Giraudel, A., Galibert, O.: The etape corpus for the evaluation of speech-based tv content processing in the french language. In: International Conference on Language Resources, Evaluation and Corpora (2012)
6. Gravier, G.and Bonastre, J., Geoffrois, E., Galliano, S., McTait, K., Choukri, K.: Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en franais. In: Proceedings Journées d'Etude sur la Parole (JEP) (2004)
7. Hakkani-Tr, D., Béchet, F., Riccardi, G., Tur, G.: Beyond asr 1-best: Using word confusion networks in spoken language understanding. Computer Speech and Language 20, 495–514 (2006)
8. Jurafsky, D., Martin, J.: Speech and language processing. Prentice Hall (2008)
9. Kripke, S.: Naming and Necessity. Semantics of Natural Language, D. Davidson and G. Harman (1972)
10. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop. pp. 249–252 (1999)
11. Marin, A., Kwiatkowski, T., Ostendorf, M., Zettlemoyer, L.: Using syntactic and confusion network structure for out-of-vocabulary word detection. In: Proceedings IEEE Spoken Language Technology Workshop (SLT). pp. 159–164 (2012)
12. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of CoNLL-2013. pp. 188–191 (2013)
13. Nowozin, S., Lampert, C.: Structured learning and prediction in computer vision. Foundations and Trends in Computer Graphics and Vision 6 (2010)
14. Palmer, D., Ostendorf, M.: Improving information extraction by modeling errors in speech recognizer output. In: Proceedings of the First International Conference on Human Language Technology Research (2001)
15. Punyakanok, V., Roth, D., Tau Yih, W., Zimak, D.: Learning and inference over constrained output. In: Proceedings of International Joint Conferences on Artificial Intelligence (2005)
16. Raymond, C.: Robust tree-structured named entities recognition from speech. In: Proceedings of International Conference on Acoustic Speech and Signal Processing, ICASSP'13 (2013)
17. Raymond, C., Fayolle, J.: Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In: Proceedings of Traitement Automatique des Langues Naturelles (2010)

18. Rosset, S., Grouin, C., Zweigenbaum, P.: Entités nommées structurées: guide d'annotation quaero. Tech. rep., LIMSI-Centre national de la recherche scientifique (2011)
19. Subramaniam, L., Roy, S., Faruquie, T., Negi, S.: A survey of types of text noise and techniques to handle noisy text. In: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data. pp. 115–122 (2009)
20. Tur, G., Deoras, A., Hakkani-Tr, D.: Semantic parsing using word confusion networks with conditional random fields. In: Proceedings of Interspeech'13. pp. 2579–2583 (2013)