Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4514–4520
Marseille, 11–16 May 2020

Where are we in Named Entity Recognition from Speech?

Antoine Caubrière[1], Sophie Rosset[2], Yannick Estève[3], Antoine Laurent[1], Emmanuel Morin[4]
[1]LIUM - Le Mans University, France, [2]LIMSI - University of Paris-Saclay, CNRS, France,
[3]LIA - Avignon University, France, [4]LS2N - University of Nantes, CNRS, France
antoine.caubriere@univ-lemans.fr, sophie.rosset@limsi.fr, yannick.esteve@univ-avignon.fr,
antoine.laurent@univ-lemans.fr, emmanuel.morin@univ-nantes.fr

Abstract

Named entity recognition (NER) from speech is usually made through a pipeline process that consists in (i) processing audio using an automatic speech recognition system (ASR) and (ii) applying a NER to the ASR outputs. The latest data available for named entity extraction from speech in French were produced during the ETAPE evaluation campaign in 2012. Since the publication of ETAPE's campaign results, major improvements were done on NER and ASR systems, especially with the development of neural approaches for both of these components. In addition, recent studies have shown the capability of End-to-End (E2E) approach for NER / SLU tasks. In this paper, we propose a study of the improvements made in speech recognition and named entity recognition for pipeline approaches. For this type of systems, we propose an original 3-pass approach. We also explore the capability of an E2E system to do structured NER. Finally, we compare the performances of ETAPE's systems (state-of-the-art systems in 2012) with the performances obtained using current technologies. The results show the interest of the E2E approach, which however remains below an updated pipeline approach.

Keywords: Named Entity Recognition, Automatic Speech Recognition, Tree-structured Named Entity, End-to-End

1. Introduction

Named entity recognition seeks to locate and classify named entity mentions in unstructured text into pre-defined categories (such as person names, organizations, locations, ...). Quaero project (Grouin et al., 2011) is at the initiative of an extended definition of named entity for French data. This extended version has a multilevel tree structure, where base entities are combined to define more complex ones. With the extended definition, named entity recognition consists in the detection, the classification and the decomposition of the entities. This new definition was used for the French evaluation campaign ETAPE (Galibert et al., 2014).

Since the ETAPE's results publication in 2012, no new work were published, to the best of our knowledge, on named entity recognition from speech for Quaero-like tree-structured French data. Tree-structured named entities can not be tackled as a simple sequence labeling task. At the time of the ETAPE campaign, state-of-the-art works focused on multiple processing steps before rebuilding a tree structure. Conditional Random Field (Lafferty et al., 2001) (CRF) are in the core of these previous sequence labeling approaches. Some approaches (Dinarelli and Rosset, 2012; Dinarelli and Rosset, 2011) used Probabilistic Context-Free Grammar (Johnson, 1998) (PCFG) in complement of CRF to implement a cascade model. CRF was trained on components information and PCFG was used to predict the whole entity tree. The ETAPE winning NER system (Raymond, 2013) only used CRF models with one model per base entity.

Most of the typical approaches for named entity recognition from speech follows a two steps pipeline, with first an ASR system and then a NER system on automatic transcriptions produced by the ASR system. In this configuration, the NER component must deal with an imperfect transcription of speech. As a result, the quality of automatic transcriptions has a major impact on NER performances (Ben Jannet et al., 2015).

In 2012, HMM-GMM implementations were still the state-of-the-art approaches for ASR technologies. Since this date, the great contribution of neural approaches for NER and ASR tasks were demonstrated.

Recent studies (Lample et al., 2016; Ma and Hovy, 2016) improve the NER accuracy by using a combination of bidirectional Long Short-Term Memory (bLSTM) and CRF layers.

Other studies (Tomashenko et al., 2016) are based on a combination of HMM and Deep Neural Network (DNN) to reach ASR state-of-the-art performances.

Lately, some E2E approaches for Named Entity Recognition from speech have been proposed in (Ghannay et al., 2018). In this work, the E2E systems will learn an alignment between audio and manual transcription enriched with NE without tree-structure. Other works use End-to-End approach to map directly speech to intent instead of map speech to word and then word to intent (Lugosch et al., 2019).

Theses works shows the growing interest in E2E approaches for this type of task.

In this paper, we propose a study of recent improvements for NER in the scope of the ETAPE campaign. We compare classical pipeline approaches with updated components and E2E approaches train with two kinds of strategy.

The first contribution of this paper is a 3-pass implementation in order to tackle tree-structured named entity recognition. This 3-pass implementation consists in splitting the tree-structured scheme of named entity annotation into 3 parts to allow classical sequential labeling of each part before rebuilding the complex structure.

The second contribution is an application of an E2E approach for tree-structured named entity recognition. It consists in training a system that learns the alignment between audio and textual transcription enriched with the structured named entity.

After a description of the Quaero named entity tree-structured task (Section 2.), we described our 3-pass im-

plementation, our state-of-the-arts for NER and ASR components and our E2E implementation (Sections 3. and 4.). Data sets (Section 5.), experimental results and analyses are presented (Section 6.) followed by a conclusion (Section 7.).

2. Task definition

This study focuses on tree-structured named entities following the Quaero guideline (Rosset et al., 2011). This guideline allows annotation according to 8 main types of named entities: *amount*, *event*, *func*, *loc*, *org*, *pers*, *prod* and *time*. The annotation uses sub-types to set up a hierarchy of named entities in order to better describe the concepts. Final annotation is necessarily the leaf of the hierarchical tree with each annotation node separates by a point. For example *loc.add.phys* which is the physical address of a place. With types and sub-types, there is 39 possible entity types in the Quaero annotation guideline.
Also, in order to decompose the concepts, named entities are annotated by component. There is 28 possible component in the Quaero annotation guideline. The component is the smallest annotated element. Each word located inside a named entity needs to be annotated in components. Except for some articles and linking words. Most of the components depend on named entities types (e.g "day", "week" which refer to the type "time") but some are cross-cutting (e.g "kind", "qualifier" which can be located inside all named entity types).
Finally, annotations have a tree-structure. A named entity can be composed of components and other named entities, itself composed of components and named entities without nesting limit. For example, the sentence "la mairie de paris" can be annotated as "la <**org.adm** <**kind** mairie > de <**loc.adm.town** <**name** paris > > >". **org.adm/loc.adm.town** are Named Entities types with sub-types and **kind/name** are components.
With the Quaero definition of named entity, NER consists in entity detection, classification and decomposition. Since this new definition is used for the French evaluation campaign ETAPE, the task in this study consists in Quaero named entity extraction from speech.

3. Pipelines systems

3.1. 3-pass implementation

Our NER systems use standard BIO2 (Sang and Veenstra, 1999) format. This standard consists of writing a column file with first the words column and then the labels column. There is one couple of word/label per line and two different sentences are separated by an empty line. The label of a word corresponds to the named entity concept in which the word is located. This label is prefixed by a "B-" or an "I-" depending on the position of the word in the concept. "B-" (Begin) is used to prefixed the label of the first word and "I-" (Inside) for all the others. "O" (Outside) is the label used for words that are not inside a concept.
Due to the structure of the annotation, most of the time words are inside more than one concept. Consequently, multiple labels are often related to a word. A single sequence labeling system cannot manage more than one prediction by word. The label concatenation can handle this problem by reducing all labels related to a word into a single one. Figure 1 illustrates an example of this concatenation.

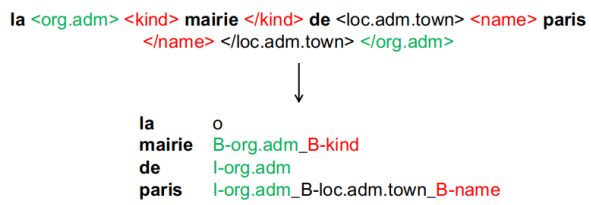

Figure 1: Transformation example of a tree-structured named entity sequence into BIO format. This sentence means in English "the town hall of paris"

The label concatenation induces a dramatic increase of the number of predictable outputs for a classical sequence labeling approach. With this concatenation this number grow up to around 1690 predictable tags. It also induces a large annotation sparsity. These issues motivated us to split the BIO annotation into different levels.
Since the named entities are necessarily decomposed in components, two facts can be deduced. First, the root of the tree structure is necessarily a named entity type. Second, the leaves of the tree structure are mainly components. And third, annotations between the leaves and the root of this structure are a mixture of type and component. Based on these observations we split the concatenated BIO annotation into three different levels.
The first level contains the furthest annotations to the word level. These annotations are the root of the tree-structured named entities. This level is represented in green color in figure 1 and requires 96 predictable tags.
The third level contains the closest annotations to the word level. These annotations are the leaves of the named entities. This level is represented in red in figure 1 and requires 57 predictable tags.
Finally, the second level contains every others annotations. These annotations are named entity types and/or components located between the root and the leaves of the named entities. This level is represented in black in figure 1 and requires 187 predictable tags.
With the annotation divided into three levels, the tree-structured NER task is tackled by three sequence-labeling systems. A sequence labeling model is trained for each level. The final output of our 3-pass implementation is the output concatenation of each model from the first level to the third. With this final output, we are able to rebuild the tree-structured annotation. Then we can transform the BIO format into sequences.
The sub-components of a named entity are dependents on the parent-component of this entity. For example, an organisation (parent-component) can contains a name (sub-component) and a time can contains an amount. In order to provide this information to our systems, the predictions from the previous levels are added as an additional input to the next levels. So, predictions from the first level are injected into the training data of the second and the third level. Also, predictions from the second level are injected

into the training data of the third level.
The 3-pass implementation is represented by the figure 2.

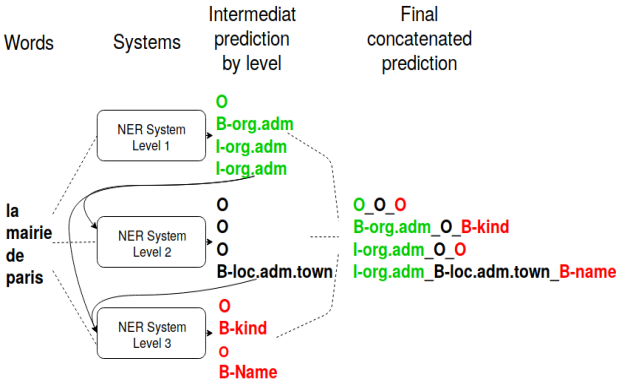

Figure 2: 3-pass implementation overview

3.2. CRF

The NER systems developed for this work are based on CRF (Conditional Random Fiedls). The models were trained using the WAPITI software (Lavergne et al., 2010). The models are based on a various set of features:

- Words and bi-grams of words located in a [-2,+2] window around the target word
- Prefixes and suffixes of words located in a [-2,+2] window around the target word
- Some *Yes/No* features like *"Does the word start with capital letter?" "Does the word contain non alphanumeric characters?"*

Some models also used morpho-syntactic features extracted from the output of the tree-tagger tool. For the 3-steps models, hypothesis provided by previous level models are also used. For all the models, we used the rprop algorithm during the training with a maximum of 40 iterations.

3.3. NeuroNlp2

NeuroNlp2[1] is an implementation of the NER system proposed in (Ma and Hovy, 2016). This system uses a neural approach for sequences labeling. It takes benefits from word and character-level embeddings learned automatically by using a combination of bidirectional Long Short-Term Memory, convolution layers and Conditional Random Fields.
A single CNN layer is used for character embeddings computing. Then, character embeddings are concatenated to word embeddings and feed the bLSTM layers. Finally, the output vectors of bLSTM are fed into the CRF layer to decode the best label sequence. In complement, dropout layers (Srivastava et al., 2014) are applied on input and output vectors of bLSTM and on input vectors of CNN.
For our works, we kept all the default parameters except the numbers of bLSTM hidden layers which is set to two and the number of units per hidden layers is set to 200.

[1]https://github.com/XuezheMax/NeuroNLP2

3.4. ASR System

The state-of-the-art speech recognition system for this study was built using Kaldi (Povey et al., 2011). The acoustic model is based on the lattice-free MMI, so-called "chain" model (Povey et al., 2016). We used a time-delay neural network (Peddinti et al., 2015) and a discriminative training on the top of it using the state-level minimum Bayes risk (sMBR) criterion (Veselỳ et al., 2013). A regular backoff n-gram model was estimated using the data presented in section 5.2. using SRILM. A 2-gram decoding is performed, followed by a 3-gram and a 4-gram rescoring step. The LM interpolation weights between the different data sources was optimized on the REPERE (Giraudel et al., 2012) development corpus. The vocabulary contains the 160k most frequents words in the manually transcribed corpus.

4. End-to-End System

In this study, we used an End-to-End (E2E) implementation based on DeepSpeech 2 ASR system (Amodei et al., 2016). His architecture consists of a stack of two 2D-invariant convolutional layers (CNN), five bidirectional long short term memory layer (bLSTM) with sequence-wise batch normalization and a final softmax layer.
This system is trained with the Connectionist Temporal Classification (CTC) loss function which allows the system to learn an alignment between an audio input and a character sequence to produce (Graves et al., 2006). Input features are sequences of log-spectrograms of power normalized audio clips calculated on 20ms windows. As we proposed in (Ghannay et al., 2018), output sequences consist of a sequence of characters composed of the word and Named entity tags. These tags are represented by starting tags and ending tags before and after words supporting these tags. The NE tree structure can be represented by a succession of tags, thus, the concatenation of labels is not required. The labels sparsity issue of the BIO format is not present in the case of our E2E system and so, the 3-pass implementation is not used. This system will learn the alignment between audio and character sequences enriched with NE tags. For example, the sentence: "la mairie de paris" for speech recognition becomes: "la <**org.adm** <**kind** mairie > de <**loc.adm.town** <**name** paris > > >" for Named Entity Recognition. In this example, "org.adm", "kind", "loc.adm.town" and "name" are four NE starting tags and ">" represent the ending tags.
Notice that starting and ending tags are actually represented by a single character within the character sequence produced by the neural network. The previous example become "la **$ &** mairie > de **% #** paris > > >".

5. Data

5.1. Named entity recognition

For our experiments, data comes from the French corpus ETAPE (Gravier et al., 2012). This corpus is composed of data recorded from French radio and TV stations between 2010 and 2011. They come from four different sources: France Inter, LCP, BFMTV, and TV8. This corpus contains 36 hours of speech divided into three parts: training

(22 hours), development (7 hours) and test (7 hours). These data have manual transcriptions and are fully manually annotated with named entities concepts.
Our training data were augmented with the Quaero corpus (Grouin et al., 2011). This corpus is composed of data recorded from French radio and TV stations between 1998 and 2004. These data are made up of 100 hours of speech manually transcribed and fully annotated with named entities following the Quaero annotation guideline.

5.2. Automatic speech recognition

In this study, we used several corpora (ESTER 1&2 (Galliano et al., 2009), REPERE (Giraudel et al., 2012) and VERA (Goryainova et al., 2014)) for a total of around 220 hours of speech. These data are used for the acoustic model training of the kaldi ASR system of the pipeline approach. The LM of this approach was trained using the speech transcripts augmented with several French newspapers (see section 4.2.3 in (Deléglise et al., 2009)). For ASR parts, our pipeline system and our E2E system use the same dataset except for the speech of ETAPE train dataset which is used only with our E2E approach.

6. Experiments

All our experiments are evaluated on the ETAPE test set with the Slot Error Rate (SER) metric (Makhoul et al., 1999) defined as:

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_b + \alpha_3 S_{bt} + \beta D + \gamma I}{R} \quad (1)$$

where:

- S_b : the number of slot boundaries substitution
- S_t : the number of slot type substitution
- S_{bt} : the number of slot boundaries and type substitution
- D/I : the number of slot deletion / insertion
- R : the number of slot references

A slot is defined as an annotated text segment with start/end boundaries and a NE type. α_1, α_2, α_3, β and γ are the weights assigned to each type of error. Here, α_3, β and γ are set to 1 and α_1 and α_2 are set to 0.5.
The best NER system of ETAPE campaign (Raymond, 2013) was made of 68 different binary CRF models. One per entity type and component. This system was applied to the output of the best ASR system and this combination reached 59.3% of SER. This constitute our baseline (System 0).
In order to use the automatic transcriptions provided by different ASR system, manual references of named entities are projected on automatic transcriptions. Also, as the E2E system produce words and NE concepts, we keep only the word to get his automatic transcriptions.
To be fully comparable, we use the ETAPE evaluation and projection scripts for all our experiments.

6.1. Pipeline Experiments

In this study, the pipeline experiments were carried out on automatic transcriptions coming from two different ASR systems. We compare the results of the best ASR of the evaluation campaign (Galibert et al., 2014) and the results of our state-of-the-art ASR. Performances of these ASR systems are presented in table 1. Evaluation Metric used is Word Error Rate (WER). The best ASR system of the evaluation campaign is denoted ASR_{2012}, while our state-of-the-art ASR is denoted ASR_{2019}. The system ASR_{2019} is trained with all our audio data described in 5.2..

Table 1: Automatic Speech Recognition performances

ASR System	WER
ASR_{2012}	21.8
ASR_{2019}	16.5

The system ASR_{2012} reached 21.8% of WER. Our state-of-the-art ASR reaches 16.5% of WER on the ETAPE test set. This represents a relative improvement of 24.3% in terms of WER.
NER systems were trained on manual transcriptions and then applied on automatic transcriptions. Part-of-speech (POS) tags were used for all of these experiments.
System *A* corresponds to a 1-pass implementation of a classical CRF approach applied on ASR_{2012}.
System *B* corresponds to a 3-pass implementation of the same CRF approach than system *1* applied on the same automatic transcription.
System *C* corresponds to the same 3-pass CRF implementation than system *B* applied on automatic transcription of our state-of-the-art ASR system.
System *D* corresponds to a 3-pass implementation of our state-of-the-art NER system applied on ASR_{2012}.
Finally, system *E* corresponds to the combination of our state-of-the-art ASR and NER systems with a 3-pass implementation of NER component.
Results of these systems are shown in Table 2.

Table 2: Pipeline experimental results

System	SER
Sys 0. Baseline ETAPE 2012	59.3
Sys A. 1-pass – CRF – ASR_{2012}	69.4
Sys B. 3-pass – CRF – ASR_{2012}	59.5
Sys C. 3-pass – CRF – ASR_{2019}	55.0
Sys D. 3-pass – bLSTM-CRF – ASR_{2012}	56.1
Sys E. 3-pass – bLSTM-CRF – ASR_{2019}	**51.1**

Our simplest system *A* reached 69.4% of SER. By using the 3-pass approach in the same configuration, the system *B* reached 59.5% of SER. The use of the 3-pass approach allows a 14.3% relative gain, showing the interest of the 3-pass approach. Results obtained with *B* are close to the baseline system (+0.2%), with only 3 CRF models instead of 68.
As expected, automatic speech transcription quality improvement has a positive impact on SER results. This can be shown by a comparison between systems *B* and *C* and

also between the systems *D* and *E*. For a CRF NER system, results start at 59.5% of SER and decrease to 55.0% (7.6% relative gain). For a bLSTM-CRF NER system, results start at 56.1% of SER and decrease to 51.1% (8.9% relative gain).
The use of our state-of-the-art NER system allows another significant improvement. This improvement can be shown by a comparison between system *B* and system *D* and also by analyzing the differences between *C* and *E*. For an HMM-GMM ASR system, results decrease from 59.5% of SER to 55.0% (7.6% relative gain) For an HMM-DNN ASR system, results decrease from 55.0% to 51.1% of SER (7.1% relative gain).
Finally, the combination of our 3-pass approach with state-of-the-art ASR and NER systems reach the best results for tree-structured named entity recognition from speech on these data at 51.1% of SER with a pipeline approach.

6.2. End-to-End Experiments

For the E2E system training, we apply the same strategy as in our previous work to compensate the lack of audio data with a manual NE annotation (Ghannay et al., 2018). It consists of multi-task learning with first an ASR system and then, by transfer learning, a NER system ($ASR -> NER_{struct}$). The output labels change between ASR and NER tasks by the addition of labels for NE tags. For the transfer learning, we keep all the model's parameters except the top layer (softmax) which are fully reset. To train the ASR task, we use all our audio data described in 5.2.. For the NER task, we use the data described in 5.1..
Our previous work shows the interest of a Curriculum-based Transfer Learning approach (CTL) for the E2E system (Caubrière et al., 2019). It consists to train the same model several times with different tasks ordered from the most generic to the most specific.
In our targeted task, a NE is composed of types and components. Components are used to decomposed NE types (see section 2.). With the CTL approach, we proposed to train the NER task with two different tasks. First with the NE types only and second with the full annotation. Since the components are directly dependent on the NE types, we assume that a task with types only is more generic than a task with types and components. We train the learning chain $ASR -> NER_{types} -> NER_{full}$, with first the speech recognition system, then the NER system trained with only the NE types annotations and finally the NER system with the full annotation for the targeted task.
Results of the both E2E systems are reported in table 3. Metrics and data sets used are the same as our pipeline experiments.

Table 3: End-to-End experimental results with a greedy decoding

System	SER
$ASR-> NER_{struct}$	62.9
$ASR-> NER_{types}-> NER_{full}$	**61.9**

Results shows the interest of the CTL approach for our task. by splitting the training into two different tasks we are able to reduce the SER from 62.9% to 61.9%.
With the DeepSpeech 2 implementation, it is possible to compute a beam search on the neural network outputs. We use two different word-level language models (3-gram and 4-gram) trained on the ETAPE and QUAERO train set. Results are presented in table 4.

Table 4: End-to-End experimental results with a beam search decoding

System	LM	SER
$ASR-> NER_{struct}$	3-gram	57.9
$ASR-> NER_{types}-> NER_{full}$	3-gram	57.5
$ASR-> NER_{struct}$	4-gram	57.3
$ASR-> NER_{types}-> NER_{full}$	4-gram	**56.9**

As expected, all results are improved by the use of a language model. By applying the 3-gram LM we can reduce significantly the SER from 62.9% to 57.9%. We can reduce more the SER by applying a 4-gram LM and reach 57.3%. Notice that the CTL approach is still useful and set our best results to 56.9% of SER.

6.3. Global comparison

We reported in table 5 the results of the best pipeline system, the best E2E system and the best system of the ETAPE campaign, our baseline.

Table 5: Reported results of ETAPE baseline and best pipeline and end-to-end systems.

System	SER
(Sys 0) Baseline ETAPE 2012	59.3
(E2E) $ASR-> NER_{types}-> NER_{full}$ (4-gram)	56.9
(PIP) 3-pass – bLSTM-CRF – ASR_{2019}	**51.1**

With our E2E approach, we reach a relative improvement of 4% since the publication of ETAPE results. However, results show also that a pipeline approach with each component updated with our 3-pass implementation still better and set the new state-of-the-art. Comparison between the baseline and our best pipeline systems shows a significant relative improvement of 13.8%. By comparison between our best E2E approach and our best pipeline approach, results show a relative improvement of 10.2% at the advantage of the pipeline approach.

7. Conclusion

This study gives an update on the NER results that can be achieved on the French ETAPE evaluation campaign. Our experiments have been carried out on pipeline and end-to-end systems. In this paper, an original 3-pass implementation is proposed for the NER component in the context of pipeline systems. By splitting the tree-structured named entities annotations into three parts, we are able to handle this task as three different simple sequence labeling tasks. This approach reaches similar results than the best NER system of ETAPE campaign with only 3 CRF models instead of 68 binaries models. Based on our previous work on flat

named entity recognition system with E2E approach, we also proposed an E2E system for structured named entity recognition. We are able to reach the best results with the E2E systems by the use of our CTL approach. By comparison between the best result of ETAPE evaluation campaign and our best E2E system, results show a relative improvement of 4%. However, this approach doesn't set the new state-of-the-art which is set by the fully updated pipeline systems with our original 3-pass implementation. Experimental results show an interesting global relative improvement of 13.8% between ETAPE results and the new state-of-the-art.

8. Acknowledgements

This work is partially supported by the French National Research Agency under grant ANR-15-CE23-0025-01 (ContentCheck project) and by the RFI Atlanstic2020 RAPACE project.

9. Bibliographical References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning*, pages 173–182.

Ben Jannet, M. A., Galibert, O., Adda-Decker, M., and Rosset, S. (2015). How to evaluate asr output for named entity recognition? In *Interspeech*, Dresden, Germany, September.

Caubrière, A., Tomashenko, N., Laurent, A., Morin, E., Camelin, N., and Estève, Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Interspeech*.

Deléglise, P., Esteve, Y., Meignier, S., and Merlin, T. (2009). Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate? In *Tenth Annual Conference of the International Speech Communication Association*.

Dinarelli, M. and Rosset, S. (2011). Models cascade for tree-structured named entity detection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Dinarelli, M. and Rosset, S. (2012). Tree representations in probabilistic models for extended named entity detection. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 174–184, Avignon, France, April.

Galibert, O., Leixa, J., Adda, G., Choukri, K., and Gravier, G. (2014). The ETAPE speech processing evaluation. In *Proc of LREC*, Reykjavik, Iceland. ELRA.

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., and Morin, E. (2018). End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107.

Goryainova, M., Grouin, C., Rosset, S., and Vasilescu, I. (2014). Morpho-syntactic study of errors from speech recognition system. In *LREC*, volume 14, pages 3050–3056.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, OR, June. Association for Computational Linguistics.

Johnson, M. (1998). Pcfg models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical Very Large Scale CRFs. In *Proceedings the 48^{th} Annual Meeting of the Association for Computational Linguistics*, pages 504–513.

Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., and Bengio, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *Proc. of DARPA Broadcast News Workshop*, pages 249–252.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glem-

bek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Raymond, C. (2013). Robust tree-structured named entities recognition from speech. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, Vancouver, Canada, May.

Rosset, S., Grouin, C., and Zweigenbaum, P. (2011). Entités nommées structurées : guide d'annotation quaero. limsi–cnrs, orsay, france.

Sang, E. F. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Tomashenko, N., Vythelingum, K., Rousseau, A., and Estève, Y. (2016). Lium asr systems for the 2016 multigenre broadcast arabic challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 285–291. IEEE.

Veselỳ, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349.