EVALUATION OF SEMANTIC ROLE LABELING AND DEPENDENCY PARSING OF AUTOMATIC SPEECH RECOGNITION OUTPUT

Benoit Favre[1,2]*, Bernd Bohnet*[1,3]*, Dilek Hakkani-Tür*[1]

[1]International Computer Science Institute, Berkeley, USA, {favre,bohnet,dilek}@icsi.berkeley.edu
[2]LIUM, Université du Maine, France, [3]University of Stuttgart, Germany

ABSTRACT

Semantic role labeling (SRL) is an important module of spoken language understanding systems. This work extends the standard evaluation metrics for joint dependency parsing and SRL of text in order to be able to handle speech recognition output with word errors and sentence segmentation errors. We propose metrics based on word alignments and bags of relations, and compare their results on the output of several SRL systems on broadcast news and conversations of the OntoNotes corpus. We evaluate and analyze the relation between the performance of the subtasks that lead to SRL, including ASR, part-of-speech tagging or sentence segmentation. The tools are made available to the community.

Index Terms— Evaluation, Semantic Role Labeling, Dependency Parsing, Automatic Speech Recognition

1. INTRODUCTION

Semantic Role Labeling (SRL), also known as Shallow Semantic Parsing, consists of extracting surface meaning in human language. These layers are identified by predicates (or frames), which evoke an action or a state, and their arguments: the agent, the patient, but also the manner, the time or the location. The availability of large text corpora annotated with semantic role labels, such as FrameNet [1], PropBank [2] and NomBank [3] allowed for development of several statistical semantic analyzers (for example [4, 5]), and shared task evaluations, such as the CoNLL 2004 and 2005. Dependency parsing is a syntactic analysis of language in which individual words of a sentence are linked to each other by head-modifier relationships (unlike constituency parsing which relies on phrase structure). Performing SRL on top of dependency trees was introduced by Hacioglu [6] and gained popularity with the 2008 and 2009 editions of the CoNLL shared tasks where it was evaluated on large text corpora [7].

On the speech processing side, SRL has become an important component of spoken language understanding systems [8, 9]. For example, ATIS dialog systems allow to obtain flight information by spotting departure and arrival information in user utterances, a shallow, task-dependent form of SRL. The MEDIA project has annotated tourist information dialogs with shallow semantics and generated a lot of interest ([10] among others). Bisazza [11], within the context of the LUNA project, studied the application of the FrameNet annotation to spoken dialogs. Recent work, such as [12, 13], only provide indirect evaluation of dependency parsing and SRL on speech, and focus on evaluating the overall understanding task. While extrinsic evaluation is useful, it does not allow researchers to assess directly the performance of their SRL components.

Performing semantic role labeling of a dependency structure is more effective for speech because head words are used to carry the information, minimizing the effect of constituent segmentation and focusing the annotation on important content words.

On text, dependency parsing is evaluated by counting matching dependencies: for a given word, the head and the label should match the reference and the hypothesis. Similarly SRL is evaluated by counting matching semantic arcs: an arc corresponds to a predicate, the word labeled as head of the argument and the argument label. The evaluation of those two tasks also considers special cases for the root of the dependency tree and predicates identity. It, however, assumes that the reference and hypothesis words and sentences being evaluated are identical, which cannot be guaranteed for errorful speech recognition and sentence segmentation output.

This problem was already tackled for constituency parsing of speech by the Sparseval tool [14]. This tool performs a word-to-word alignment of the reference and the hypothesis prior to scoring parse trees. We extend this idea for evaluating the performance of the output of joint (or sequential) dependency parsing and semantic role labeling systems (Section 2). We study the evaluation of a set of SRL systems on the OntoNotes corpus (Section 3), and compare the evaluation of alignments with a metric that does not require alignment (Section 4).

2. EVALUATION METHODOLOGY

2.1. CoNLL evaluation on text

The CoNLL 2009 shared task (as well as previous instances) evaluates the performance of semantic role labeling and dependency parsing systems on a given test set by computing the recall, precision and F_1-measure of matched arcs in the semantic and syntactic trees. Recall (R) is the ratio of correct arcs ($\#corr$) in the system output compared to the reference ($\#ref$), $R = \#corr/\#ref$. Precision (P) is the proportion of correct arcs in the system output ($\#hyp$), $P = \#corr/\#hyp$. F_1-score (referred to as F-score or F-measure) is the harmonic mean between recall and precision, $F = 2PR/(P + R)$.

The semantic annotation consists of predicates and their arguments. For a given argument, the head-word of the constituent it spans (the argument-bearing word) is linked to the predicate and labeled with an argument type. The predicates are attached to the words they are evoked by, and labeled by the sense number of those words. As depicted in the top part of Figure 1 for the PropBank annotation, the verb "to affect" is a predicate and "perception" is one

We thank Richard Johansson, Anders Bjrkelund, and Sameer Pradhan for providing their SRL systems, and Wen Wang for providing ASR output. This work is supported by the Defense Advanced Research Projects Agency (DARPA) GALE project, under Contract No. HR0011-06-C-0023, by the German Academic Exchange Service (DAAD), and the Agence Nationale de la Recherche (ANR) through the PORT-MEDIA project under contract number ANR-08-CORD-026. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies.

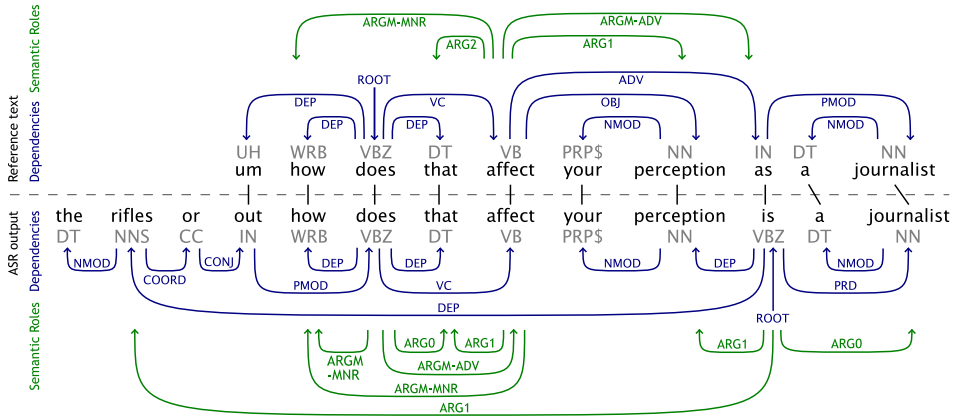

Fig. 1. Example of alignment between gold-standard data (top) and system output on speech (bottom) for dependency parsing (inner) and SRL (outer). Arrows point from the head to the dependent.

of its arguments, representing the affected entity. The syntactic tree consists of arcs between pairs of words, denoting dependency labels between words and their head (including a root arc). In the example of Figure 1, "perception" is the head of "your" with label "NMOD" which means that "your" modifies the noun "perception".

The CoNLL evaluation script[1] computes and outputs F-scores for various measures:

- dependency parsing: labeled (for a given word, the head and the label should match), unlabeled (ignores relation label), labels (ignores the head), and exact sentences (counting reference sentences).
- semantic role labeling: labeled (considers the argument label), unlabeled, propositions (a predicate and its arguments shall be correct) and exact sentences.
- Joint evaluation: the shared task intends to foster joint syntactic and semantic parsing research by using a macro average of both scores for labeled and unlabeled (with a trade-off factor of 0.5), and a micro average for which the recall and precision values are recomputed by summing $\#corr$, $\#hyp$ and $\#ref$ from both tasks (this puts more weight towards dependency parsing because it has more arcs).

2.2. Extensions for speech

Automatic speech recognition (ASR) output tends to contain errors, due to difficult acoustic conditions and mismatches between models and test conditions among others. These errors result in inserted, deleted and substituted words, which prevent the use of the CoNLL tools that assume the same word string in both reference and hypothesis for evaluation. In addition, automatic sentence segmentation may result in errors like split or merged sentences which confuse the scoring process as reference and hypothesis trees do not match anymore. We propose to follow the work by Roark [14] and perform a show-level alignment between the hypothesis words and the reference words prior to scoring. This alignment is established by computing the Levenshtein minimum edit distance found through dynamic programming in $O(n^2)$.

Then, a dependency arc is considered correct if the head of a word in the reference is aligned to the head of the word it is aligned to in the hypothesis (see Figure 2). Similarly, a semantic arc is considered correct if the word bearing the argument of a predicate in the reference is aligned to the argument of the corresponding predicate in the hypothesis. Note that virtual arcs are used to represent the root of the dependency tree, and the sense of the predicates. In this setting, a dependency (or SRL) arc can be considered correct even if words do not match (but they must be aligned), which decreases the correlation with word error rate. Another approach would be to transfer the reference annotation to the ASR output, but it would not be easy to assess inserted and deleted words (e.g., if the word supporting an argument is deleted, should the reference argument be placed on the previous word?)

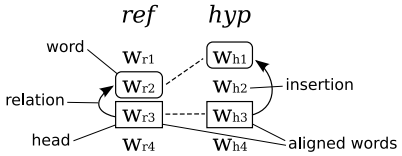

Fig. 2. Conditions for an arc to be correct for dependency parses.

We extend the CoNLL'09 evaluation metrics with the new set of matching conditions on aligned reference and hypothesis. Because the metrics do not rely on word identity directly, we also add a "lexicalized" metric for which reference and hypothesis words involved in an arc must match. As a by-product of the alignment, our evaluation tool also outputs word error rate, sentence boundary detection F-score (a boundary between two consecutive words with alignments should result in a boundary between the two aligned words in the hypothesis), and part-of-speech tagging F-score. An example of the complete output is given in Figure 3. Note that in the case of identical words, our tools yield the same results as the CoNLL eval script.

In addition, we compute the "bag-of-relation" syntactic and semantic F-score for which ordering and location of words is ignored. A bag of relation is formed from triples (word, head-word, label) for syntax and (argument-word, predicate-word, argument-label) for semantics, for the reference and the hypothesis. For a given relation, the number of correct is the minimum of its number of occurrences in the reference and the hypothesis, leading to the number of correct relations:

$$\#corr = \sum_{rel} min\left\{\#hyp(rel), \#ref(rel)\right\}$$

[1] http://ufal.mff.cuni.cz/conll2009-st/scorer.html

Such an evaluation metric is interesting because it does not require word-to-word alignment but it may be affected by word error rate, and by factors related to word order (for example, it gives the same score if the whole input is presented backwards). Our evaluation tools are made available[2] to the community so that future research can use a common evaluation framework.

Category	Type	R	P	F
Syntax	labeled	67.27	69.34	68.29
	unlabeled	70.48	72.65	71.55
	labels	72.42	74.65	73.52
	lexicalized	64.94	66.94	65.93
	exact sentences	9.68	11.24	10.40
	bag-of-relation	67.71	69.80	68.74
Semantics	labeled	63.41	65.07	64.23
	unlabeled	71.04	72.91	71.96
	lexicalized	61.49	63.10	62.28
	predicates	87.94	87.49	87.72
	propositions	25.12	24.99	25.06
	exact sentences	21.25	24.66	22.83
	bag-of-relation	65.01	66.72	65.86
Joint	macro labeled (0.5)	65.34	67.21	66.26
	macro unlabeled (0.5)	70.76	72.78	71.76
	micro labeled	65.98	67.91	66.93
	micro unlabeled	70.67	72.74	71.69
Other	Word error rate (%)	16.01		
	Sentence boundaries	48.53	56.33	52.14
	Part-of-speech tags	83.89	86.48	85.17

Fig. 3. Example output of the evaluation tool (LUND system on ASR output).

3. DATA AND SYSTEMS

In the following experiments, we use the OntoNotes release 3 data which covers English news and conversation broadcasts [15]. The data is annotated with constituency trees and PropBank-style semantic frames (verbs only). We obtained speech recognition output from SRI's broadcast news recognizer [16] for a subset of the BN and BC files and split this set for training and testing.[3] Since the reference transcripts are based on closed captions and lack timing information, some shows yield very high word error rates (WER) with long strands of unmatched words (untranscribed commercials, for example). Therefore, the shows with a WER higher than 50% have not been included in the test set. The resulting average WER is 16%. Sentence segmentation of the ASR output was realized with a simplified system which classifies each inter-word as sentence boundary or not using lexical features and pause duration (performing at a F-score of 52.14). Measured on reference text, the training set consists of 242,814 words (16,076 sentences) and the test set contains 27,393 words (1,807 sentences). We do not intend to tune the systems, therefore no development set is defined. The test set contains 3,873 verb predicates (auxiliaries are ignored) which bear 9,875 arguments labeled with 20 classes.

Constituency trees are converted to the dependency formalism using the pennconverter tool [17], which was used for preparing data in the CoNLL shared tasks. Then, the semantic annotation is transfered to the head-words of the phrase-level annotation.

[2]http://code.google.com/p/srleval

[3]Test set: bn/voa_0246 to 264, bn/pri_0106 to 112, bn/nbc_0035 to 39, bn/mnb_0024 and 25, bn/cnn_0381, 400, 422 to 424 and 431, bn/abc_0066 to 69, bc/cnn_0008, bc/msnbc_0007

We evaluated four systems which performed very well in at least one of the previous iterations of the CoNLL shared task. The systems are the MATE system [18], the LUND system [19], the LTH system [20] and the ASSERT system [21]. They have been retrained on the speech transcripts when possible to minimize the mismatch with ASR output (punctuation removed, lower case, numbers converted to words, tokenization). The MATE system, which participated to the CoNLL'09 evaluation, contains all components necessary for semantic role labeling of speech: part-of-speech tagging, lemmatization, dependency parsing, predicate identification, argument identification, argument labeling and word sense disambiguation. We used the MATE processing chain for providing dependency parses to the LUND and LTH systems (from which we only used the SRL modules). The ASSERT system relies on constituency parses which we provided using the Berkeley Parser [22], retrained on speech. Its output was converted to the dependency formalism for scoring. We could not, however, retrain ASSERT's SRL module and used the stock model trained on text. Since none of the systems was actually tuned for processing speech (merely retrained on matching data), one should not draw conclusions regarding which is better.

4. RESULTS AND ANALYSIS

In terms of dependency parsing, we have only one system to evaluate, the MATE system. It performs at 84.90 (labeled F-score) on reference text (no punctuation, lower case), 73.21 on ASR output with reference sentence segmentation, and 68.29 on the full automatic pipeline. The decrease in performance seems to be correlated with part-of-speech tagging as evidenced in Figure 4.

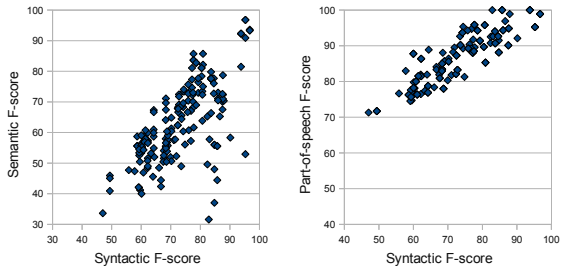

Fig. 4. Relation between "show-level" syntactic and semantic scores (left), and syntactic and part-of-speech scores (right).

Figure 5 displays SRL labeled F-scores for all the systems on three conditions: reference parses (ref dep), reference text but automatic parsing (text), ASR output with reference sentence segmentation (ASR-rs) and ASR full automatic (ASR). The performance of all systems seem to loose around 10 points by going to automatic parsing and 10 points by processing ASR output. Sentence segmentation claims another 3 points even though about 50% of the sentences are wrong. Figure 4 shows that, with the exception of a few outliers, semantic and syntactic F-scores are quite correlated at the show level. The outliers have been verified to be small shows with labeling errors in the reference (e.g. inverted passive).

We compare the word-aligned evaluation to the bag-of-relation metric in Figure 6. The two are quite correlated (0.93-0.96) which validates the bag-of-relation approach for skipping the word-to-word alignment process which does not scale well and can lead to problems when, for instance, two words in the ASR output shall be aligned to one word in the reference: The assignment is arbitrary which deteriorates the evaluation when one of the words bears a

System	Ref dep	Text	ASR-rs	ASR
LUND	86.76	76.79	67.81	64.23
MATE	82.57	73.20	64.89	61.62
LTH	79.41	72.07	64.05	60.80
ASSERT	n/a	58.67	52.23	50.76

Fig. 5. Labeled semantic F-measure for SRL only (on reference parses), full pipeline on text, ASR output with reference sentences (ASR-rs) and ASR output with automatic sentences.

relation. On the other hand, the "bag-of-relation" metric is more affected by word errors since a relation is considered wrong if the words do not match.

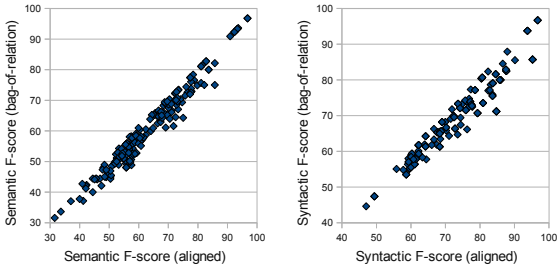

Fig. 6. Word-aligned evaluation compared to the location unaware bag-of-relation evaluation. The correlation coefficient is $R^2 = 0.96$ for semantics and $R^2 = 0.93$ for syntax.

Figure 7 shows the relation between word error rate (WER) and syntactic and semantic F-scores. The loss is about 3 points of semantic F-score per point of WER and 1.5 point of syntactic F-score for each point of WER. The same outliers as in Figure 4 are observed in this plot.

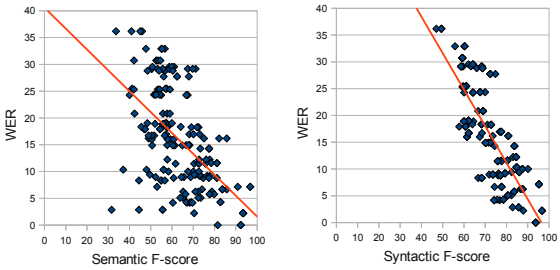

Fig. 7. Plot of "show-level" semantic and syntactic F-score against word error rate. Only one system is shown for syntax, the aggregate of all systems for SRL.

5. CONCLUSIONS

We have adapted semantic role labeling and dependency parsing evaluation to account for word errors and sentence segmentation mismatch in ASR output. An analysis of the performance of retrained systems shows that the errors of at different levels (part-of-speech tagging, dependency parsing and SRL) are quite correlated. The likely source of this observation is that the systems are trained on reference data. One way of improving SRL on speech could be to retrain systems on ASR output or modify them to process word lattices. Another interesting extension would be to evaluate SRL of machine translation output, where the words will be even further from the reference text. We believe that the public availability of an evaluation toolkit will accelerate research on SRL for speech data.

6. REFERENCES

[1] C. Baker, C. Fillmore, and J. Lowe, "The Berkeley FrameNet Project," in *International Conference on Computational Linguistics*, 1998, pp. 86–90.

[2] P. Kingsbury and M. Palmer, "From treebank to propbank," in *LREC*, 2002, pp. 1989–1993.

[3] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman, "The NomBank project: An interim report," in *HLT-NAACL*, 2004, pp. 24–31.

[4] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Computational Linguistics*, vol. 28, no. 3, pp. 245–288, 2002.

[5] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky, "Shallow semantic parsing using support vector machines," in *HLT-NAACL*, 2004.

[6] K. Hacioglu, "Semantic role labeling using dependency trees," in *International Conference on Computational Linguistics*, 2004.

[7] J. Hajic, M. Ciaramita, R. Johansson, D. Kawahara, MA Martı, L. Marquez, A. Meyers, J. Nivre, S. Pado, J. Stepanek, et al., "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages," in *CoNLL*, 2009, vol. 5.

[8] N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert, "The AT&T spoken language understanding system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 213–222, 2006.

[9] R. De Mori, F. Bechet, D. Hakkani-Tür, M. McTear, G. Riccardi, and G. Tur, "Spoken Language Understanding for Conversational Systems," *SPM Special Issue on Spoken Language Technologies*, vol. 25, no. 3, pp. 50–58, May 2008.

[10] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *European Conference on Speech Communication and Technology*, 2005.

[11] A. Bisazza, M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Semantic annotations for conversational speech: from speech transcriptions to predicate argument structures," in *SLT*, 2008.

[12] F. Bechet and A. Nasr, "Robust Dependency Parsing for Spoken Language Understanding of Spontaneous Speech," in *Interspeech*, 2009, pp. 1039–1042.

[13] C.-H. Liu and C.-H. Wu, "Semantic role labeling with discriminative feature selection for spoken language understanding," in *Interspeech*, 2009, pp. 1043–1046.

[14] B. Roark, M. Harper, E. Charniak, B. Dorr, M. Johnson, J. Kahn, Y. Liu, M. Ostendorf, J. Hale, A. Krasnyanskaya, et al., "Sparseval: Evaluation metrics for parsing speech," in *LREC*, 2006.

[15] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: the 90% solution," in *NAACL*, 2006, pp. 57–60.

[16] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde, and J. Zheng, "SRIs 2004 broadcast news speech to text system," in *EARS Rich Transcription Workshop*, 2004.

[17] R. Johansson and P. Nugues, "Extended constituent-to-dependency conversion for English," in *NODALIDA*, 2007.

[18] B. Bohnet, "Efficient Parsing of Syntactic and Semantic Dependency Structures," in *CoNLL*, 2009, pp. 67–72.

[19] A. Bjorkelund, L. Hafdell, and P. Nugues, "Multilingual Semantic Role Labeling," in *CoNLL*, 2009, p. 43.

[20] R. Johansson and P. Nugues, "Dependency-based semantic role labeling of PropBank," in *EMNLP*, 2008, pp. 69–78.

[21] S. Pradhan, K. Hacioglu, W. Ward, J.H. Martin, and D. Jurafsky, "Semantic role chunking combining complementary syntactic views," in *CoNLL*, 2005, p. 217.

[22] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *NAACL*, 2007, pp. 404–411.