Selection and Aggregation Techniques for Crowdsourced Semantic Annotation Task

Shammur Absar Chowdhury[1], *Marcos Calvo*[2], *Arindam Ghosh*[1], *Evgeny A. Stepanov*[1], *Ali Orkan Bayer*[1], *Giuseppe Riccardi*[1], *Fernando García*[2], *Emilio Sanchis* [2]

[1]Dept. of Information Engineering & Computer Science, Univ. of Trento, Trento, Italy
[2] Dept. de Sistemes Informàtics i Computació, Univ. Politècnica de València, València, Spain
{sachowdhury,aghosh,stepanov,bayer,riccardi}@disi.unitn.it,
{mcalvo,fgarcia,esanchis}@dsic.upv.es

Abstract

Crowdsourcing is an accessible and cost-effective alternative to traditional methods of collecting and annotating data. The application of crowdsourcing to simple tasks has been well investigated. However, complex tasks like semantic annotation transfer require workers to take simultaneous decisions on chunk segmentation and labeling while acquiring on-the-go domain-specific knowledge. The increased task complexity may generate low judgment agreement and/or poor performance. The goal of this paper is to cope with these crowdsourcing requirements with *semantic priming* and unsupervised quality control mechanisms. We aim at an automatic quality control that takes into account different levels of workers' expertise and annotation task performance. We investigate the judgment selection and aggregation techniques on the task of cross-language semantic annotation transfer. We propose stochastic modeling techniques to estimate the task performance of a worker on a particular judgment with respect to the whole worker group. These estimates are used for the selection of the best judgments as well as weighted consensus-based annotation aggregation. We demonstrate that the technique is useful for increasing the quality of collected annotations.

Index Terms: Crowdsourcing, Annotation, Cross-language porting

1. Introduction

In recent years crowdsourcing has been successfully applied to a variety of research problems. It has been widely used to perform tasks where it is difficult, expensive or time consuming to find enough experts. The main crowdsourcing paradigm involves breaking down complex tasks into smaller components (micro-tasks) before distributing them to the crowd. In the Natural Language Processing (NLP) domain, off-the-shelf crowdsourcing platforms have been successfully used for tasks like corpus creation [1, 2], transcription [3, 4], translation [5], and annotation [6, 7]. In this paper we address the task of cross-language transfer of semantic annotation. Particular properties of the task are its complexity, since workers must make simultaneous decisions on both segmentation and labeling while acquiring domain specific knowledge on-the-go; and lack of reference annotations in the language of the task, which makes the worker quality control difficult.

Since the attention span of crowd workers in crowdsourcing platforms is short – they promote speed, not accuracy – using these platforms for such complex tasks requires very careful task design. Since workers have different levels of expertise and task performance, unrestricted annotation introduces noise, and the quality of the collected data will vary.

Researchers have experimented with several techniques to improve the quality of crowdsourced annotations. The first step to generate high quality annotations is to use quality control mechanisms, such as qualification tasks and gold standard references, to filter out low quality crowd-workers. Researchers have also shown that crowds can be "taught" to perform better while carrying out the task. In [8], the authors successfully used motivational feedback as a training signal to improve workers' performance.

Targeted crowdsourcing has recently evolved as another paradigm to attract high quality workers. Such platforms can be designed to effectively avoid spammers, and target users possessing specialized skill sets (domain knowledge and language skills).

Since crowdsourcing tasks yield multiple annotations for each sub-task, there is a need for aggregation and consolidation of these multiple results to arrive at a single high quality annotation. Annotation tasks are of different complexity levels, and require different annotation selection and aggregation techniques. The task of final annotation generation can either be selecting the best judgment – annotation hypothesis – from the crowd-generated set with respect to some evaluation criteria [6], or aggregating the annotations from different crowd workers using weighted or unweighted voting schemes. In [9] it was demonstrated that individual worker judgments can be weighted using their performance on the gold standard examples to correct for the individual biases. In [10] the authors have extended these methods and proposed to weight worker judgments based on their agreement with the full worker population.

We propose unsupervised stochastic modeling techniques for the selection of the best user judgments and their consensus-based aggregation. Specifically, for the task of cross-language semantic annotation transfer, where the goal is to transfer concept annotation from one language to another, we use joint language models, trained on word-concept language pairs from worker judgments, to score individual annotation hypotheses.

The paper is structured as follows. In Section 2 we describe our semantic annotation transfer model, the task design, and the stochastic modeling techniques for automatic annotation selection and consensus based aggregation. In Section 3 we evaluate the proposed techniques on the data collected through *targeted* crowdsourcing. Section 4 summarizes the results and provides concluding remarks.

2. Methodology

In a typical annotation task a set of items U (e.g. utterances, images, etc.) is annotated by a set of annotators A to yield a set of annotation hypotheses H, such that:

$$U = \{u_1, ..., u_i, ..., u_n\}$$
$$A = \{a_1, ..., a_j, ..., a_m\}$$
$$H = U \times A = \{h_{1,1}, ..., h_{n,m}\}$$

The matrix H is a sparse one, since each utterance u_i is annotated only by a subset of annotators A_i. Let $H_{i,*}$ represent a set of annotation hypotheses for an utterance u_i (row in the matrix H), and $H_{*,j}$ represent a set of annotation hypotheses by annotator a_j (column in the matrix H), such that:

$$H_{i,*} = \{h_{i,1}, ..., h_{i,m}\}$$
$$H_{*,j} = \{h_{1,j}, ..., h_{n,j}\}$$

An item-level annotation hypothesis $h_{i,j}$ is essentially a mapping $m_{i,j}$ selected by an annotator a_j for an item u_i from a set of all possible mappings M_i.

$$M_i = u_i \times L = \{m_{i,1}, ..., m_{i,x}\}$$
$$L = \{l_1, ..., l_x\}$$

where L is a finite set of task specific labels.

In case of semantic annotation task, where an utterance is annotated with a set of domain-specific concepts such that a concept covers a certain span of the utterance, there is one label per word. Thus, an annotation hypothesis $h_{i,j}$ is a mapping $m_{i,j}$, which itself is a mapping between a sequence of words W_i and a set of concepts C_j selected by annotator a_j from a set of domain concept C for the words in an utterance u_i. Thus, the set M_i of all possible mapping is more complex.

$$M_i = W_i \times C = \{m_{i,1,1}, ..., m_{i,k,l}\}$$
$$W_i = \{w_{i,1}, ..., w_{i,k}\}$$
$$C = \{c_1, ..., c_l\}$$

2.1. Crowdsourced Annotation Task Design

The goal of cross-language semantic annotation transfer task is to generate an annotation in the target language which is as much as possible close to the source language annotation. The ultimate goal of the annotation is to train machine learning algorithms. The most important factor for machine learning is consistency of the annotations. Thus, we want crowdsourced annotations to be consistent within themselves and with the source language annotation. Since concept annotations in the source language are domain-specific, either the task has to be simplified or the domain knowledge has to be transferred on-the-go to the annotators.

For the simplification of the annotation task one option is to reduce the label set C to more coarse-grained concept labels (model-reducing simplification [11]). The simplification is not applicable in our setting since we are loosing consistency with the source language annotation. A model-preserving alternative is to decompose the task into smaller sub-tasks (as small as pair-wise similarity judgments [11], for instance). But, this simplification would require a lot more judgments to be collected. Thus, the ideal choice, for the task, is to transfer the domain knowledge.

With respect to the annotation model we have just defined, the goal of transferring the domain knowledge is to limit the number of word-to-concept mappings $m_{i,j}$ an annotator can choose from M_i – a set of all possible mapping for the utterance u_i. Since we have the source language expert annotations, the first choice would be to allow only concepts from the source language annotation; however, such a restriction would potentially disallow concepts that otherwise the crowd would agree upon. Thus, the task is designed for *priming* the annotators with the unique list of concepts from the source language. Annotators are free to use it or ignore it altogether. To assess the utility of *priming* for the transfer of domain knowledge, we also designed a *non-primed* task. The comparison of *primed* vs. *non-primed* annotation results is provided in Section 3.3.

2.2. Annotation Selection and Aggregation

The mapping $m_{i,j}$ an annotator has provided is not necessarily the correct one. For the consistency with the source language annotation, for the utterance u_i, our goal is either to (1) **select** the mapping m'_i from the set of available mappings M_i that is the closest, or (2) to **aggregate** the available word-to-concept mappings in M_i to generate a single one that best represents the meaning of u_i.

In the absence of any other information about annotators or utterances, the baseline case for the selection would be to *randomly* pick one of the mappings from M_i. The baseline case for the aggregation, on the other hand, is for each word-to-concept mapping in M_i to select the most frequent (w_k, c_l) pairs, i.e. to use *majority voting*, and randomly or heuristically break the ties. Recognition Output Voting Error Reduction (ROVER) is one of the most frequently used tools in Automatic Speech Recognition community for the aggregation of outputs of multiple recognition systems and selection of the best scoring sequence.

For the illustration, consider a set of three annotation hypotheses for an utterance in Figure 1. The selection baseline is to randomly pick either **A1**, **A2**, or **A3**. The frequency-based aggregation baseline (**MV** row) is different from all the annotation hypotheses. The tie in this case was broken by selecting the first option.

In crowdsourcing annotators have different levels of expertise and task performance; thus, random selection and equal weighting of all the annotation hypotheses is not a good choice. While in crowdsourcing majority voting is the most common technique for hypotheses aggregation, a number of techniques were proposed in [10] to estimate the reliability of a particular annotator based on the *agreement* with the the pool of all the annotators. In the case of an item level annotation (e.g. label per utterance) considered by the authors, given the sufficient amount of annotations per item this estimate is straightforward. In the case of a few word-level annotations (3 judgments per utterance in our case) such an agreement would not be reliable. Since word-to-concept mappings are context dependent, using the word-concept pairs appearing in other utterances for agreement estimation is not an option. Moreover, since frequency-based aggregation takes local decisions for each item without taking into account the context in which it appears.

In case gold standard is available, the frequent technique for estimating the reliability of a annotator is maximum likelihood [9], which, unfortunately, requires a large amount of data per annotator. Simplifications to cope with this limitation have been proposed in [10]. The technique that uses maximum likelihood estimation and context for word-to-concept mappings is Language Models (LM). Among other tasks, LMs were already successfully applied for the task of sentence aggregation in Ma-

	Mario	*Rossi*	*trabajo*	*para*	*el*	*CSI*	*Piemonte*
A1	User.name	User.name	null	null	null	Inst.name	Inst.name
A2	User.name	User.surname	null	null	null	Inst.name	Inst.location
A3	User.name	User.surname	Action	Action	null	Inst.name	null
MV	*User.name*	*User.surname*	*null*	*null*	*null*	*Inst.name*	***Inst.name***

Figure 1: A set of three annotation hypotheses for an utterance *"Mario Rossi, trabajo para el CSI Piemonte"* ("Mario Rossi, I work for CSI Piemonte"). **MV** is the majority-voted (frequency-based) aggregation baseline with the tie for *Piemonte* is broken by selecting the first option. (Concepts are simplified and abbreviated due to space considerations.)

chine Translation [12] and Multilingual Spoken Language Understanding [13]. Thus, the relative likelihoods of the annotations with respect to the LMs trained on the crowdsourced data can be used as an annotator or annotation *agreement* estimations. Since we are dealing with word-to-concept mappings, the LM is trained on the word-concept pairs (i.e. joint LM).

Using all the collected annotations $H = \{h_{i,j}, ..., h_{n,m}\}$, we can score each annotation hypothesis under the following settings:

1. Per *utterance*: using language model trained on all the judgments except the ones for the utterance u_i, i.e. on the set of annotation hypotheses $H - H_{i,*}$; i.e. Leave-One-Utterance-Out (LOUO) setting.
2. Per *annotator*: using language model trained on all the judgments except the ones for the annotator a_j, i.e. on the set of annotation hypotheses $H - H_{*,j}$; i.e. Leave-One-Annotator-Out (LOAO) setting.
3. Per *utterance & all annotators for it*: using language model trained on all the judgments except ones by the annotators a_j in the set A_i who have annotated the utterance u_i, i.e. on the set of annotation hypotheses $H - H_{i,j}; j \in A_i$; i.e. Leave-All-Annotators-Out (LAAO) for a given utterance setting.

By averaging the per-annotation LM scores for settings 2 and 3, that give likelihoods of the mappings, we can estimate the overall agreement of an annotator. Let's call these settings LOAO-agreement and LAAO-agreement, respectively. These per-annotation LM scores can be used for the **selection** of the best hypothesis.

All these settings are *unsupervised*, we use only the crowdsourced target language annotations. Thus, they can be applied for the data collection in absence of gold standard data. The setting 1 is applicable only as post-processing of the collected data. Settings 2 and 3 are applicable in online data collection setting as well; and, after collecting several annotations, can be used as quality control thresholds.

The notions of reliability and agreement of annotators are critical for the crowdsourcing platforms, such as Amazon Mechanical Turk. In our case of *targeted* crowdsourcing, on the other hand, the crowd is already assumed to possess the desired skills and not to contain spammers; thus, the utility of the scoring should be lower. It only applies to the levels of expertise. In Section 3 we evaluate the hypothesis ranking with LM using the three settings we defined. We evaluate both hypothesis selection and aggregation.

3. Experiments and Results

This Section first describes the original corpus used for the annotation and the data collected using *targeted* crowdsourcing. Then we describe the evaluation methodology and the experiments on *priming* and the selection and aggregation techniques for semantic annotation transfer task.

3.1. Data Set

The crowdsourced semantic annotation transfer task has been done using Spanish utterances of Multilingual LUNA Corpus [14, 15]. The corpus is the professional translation of Italian LUNA Corpus [16] to Spanish, Turkish and Greek. The Italian LUNA Corpus is a collection of 723 human-machine dialogs (approximately 4K user turns) in the hardware/software help desk domain. The LUNA concept ontology – containing a total of 45 unique concepts – is arranged in a two-level hierarchy with 26 top-level concepts. The goal of crowdsourced annotation task is to transfer the concept attribute-value annotation (used for training Spoken Language Understanding models) from Italian to Spanish.

Using the *targeted* crowdsourcing platform of [17], a subset of 800 utterances was assigned to separate crowdsourcing tasks, where each task contained 50 utterances presented on 5 pages (10 utterances per page). For a period of two weeks, around fifty workers completed over 2000 primed and non-primed semantic annotation tasks. In total, we have collected 763 utterances with at least 3 annotations in primed setting and 420 utterances in non-primed setting. The effect of priming is evaluated using a common subset of 420 utterances and selection and aggregation techniques are evaluated using the 763 utterances from the primed setting.

3.2. Evaluation Methodology

Since our task is cross-language transfer of semantic annotation, we perform two-way evaluation. The consistency of collected annotations is evaluated as inter-annotator agreement using pair-wise precision, recall and F-measures, randomly assigning each annotation hypothesis one of the 3 'hypothetical' annotators (since we have collected 3 annotation hypotheses/judgments per utterance).

The cross-language transfer is evaluated against the source language references. Even though Italian and Spanish are close languages, to overcome any concept re-ordering issues due to translations and worker differences in concept segmentation, the consecutive concepts of the same type in both the source language reference and the hypotheses are first merged and the resulting list is sorted alphabetically. The two lists are aligned with respect to Levenshtein distance and precision (p), recall (r), and F-measure (F_1) are computed with respect to the alignment errors: insertions (I), deletions (D), and substitutions (S) according to the equations 1.

In the equations C is the number of correct labels. Substitution counts both in precision and recall since it can be decomposed as insertion and deletion.

$$p = \frac{C}{C+I+S}; \quad r = \frac{C}{C+D+S}; \quad F_1 = \frac{2*p*r}{p+r} \quad (1)$$

	P	R	F1
Non-Primed	36.91	34.12	35.36
Primed	62.18	55.98	**58.92**

Table 1: Inter-annotator agreement for *primed* and *non-primed* annotation settings reported as averages of pair-wise precision (P), recall (R) and F-measures (F1) for the lists of unique concepts regardless of the order.

	P	R	F1
Non-Primed	42.10	23.76	30.38
Primed	77.30	47.70	**58.96**

Table 2: Cross-Language Transfer for *primed* and *non-primed* annotation settings using random re-sampling as precision (P), recall (R) and F-measure (F1); for random re-sampling the results are averages of 1,000 iterations.

3.3. Primed vs. Non-Primed Annotation

As it was mentioned, the goal of priming is two-fold: to transfer the domain knowledge and to constrain the word-to-concept mapping choices of the crowd. Thus, we expect that the annotation hypotheses collected in primed setting will have higher inter-annotator agreement as well as will be more consistent with the source language concepts. The inter-annotator agreement for both settings are given in Table 1 and the cross-language transfer performances using random re-sampling are given in Table 2. In both cases the annotations collected using priming have much higher F-measures. Thus, we conclude that priming is effective for both domain knowledge transfer and restricting the mapping choices.

3.4. Hypothesis Selection

In this Section we evaluate our hypothesis selection methodology – scoring the hypotheses with the 3-gram LM trained under the three settings: Leave-One-Utterance-Out (LOUO), Leave-One-Annotator-Out (LOAO), and Leave-All-Annotators-Out (LAAO), and per-annotator agreements computed for the last two settings (LOAO-agreement and LAAO-agreement). As it was mentioned, the baseline of selection is random choice, and we report it as averages of random re-sampling for 1,000 iterations. The upper bound of the selection, on the other hand, is choosing the annotation hypothesis that is the closest to the source language references, i.e. *Oracle*. The results are given in Table 3. While all the settings are above the baseline, the Leave-One-Annotator-Out yields the best performance, ≈ 3.5 points higher than the baseline.

3.5. Hypothesis Aggregation

Similar to hypothesis selection with respect to the LM score, the hypotheses could be aggregated using ROVER and the utterance scores, or annotator-agreement scores we have defined. As a baseline we are using majority voted ROVER. Since the recall of the selection baseline is low, to compensate it, we compute *maximal* ROVER. The difference from *majority* ROVER is in treatment of *null* concepts, the technique ignores majority vote for *null* and accepts any concept. It is expected to increase the recall, but lower the precision.

The results for hypothesis aggregation are presented in Table 4. Majority-voted ROVER is rather a strong baseline, and it outperforms all the scoring techniques except Leave-One-Utterance-Out. Maximal ROVER, on the other hand, increases the recall by ≈ 7.7, but the precision drops by ≈ 13.5. The best performing weighting scheme, Leave-One-Utterance-Out, increases the recall by ≈ 1.2 and lowers the precision by ≈ 1.4; however, the F-measure is increased by ≈ 0.4.

	P	R	F1
Baseline: Rand. Re-sampling	81.15	55.01	65.57
Oracle	93.46	71.98	81.33
LOUO	84.75	55.34	66.96
LOAO	87.60	56.95	**69.02**
LAAO	84.45	55.49	66.97
LOAO-agreement	84.88	56.51	67.85
LAAO-agreement	84.52	57.35	68.33

Table 3: Precision (P), recall (R) and F-measure (F1) of the annotation selection using LM scores for Leave-One-Utterance-Out (*LOUO*), Leave-One-Annotator-Out (*LOAO*), Leave-All-Annotators-Out (*LAAO*), and per-annotator agreements (*LOAO-agreement* and *LAAO-agreement*). Baseline is an average of 1,000 iterations of random re-sampling. *Oracle* is an annotation selected with respect to the source language references.

	P	R	F1
ROVER: Majority	85.45	59.36	70.05
ROVER: Maximal	71.99	67.06	69.44
ROVER: LOUO	84.06	60.60	**70.43**
ROVER: LOAO	83.02	59.39	69.25
ROVER: LAAO	84.18	59.58	69.77

Table 4: Precision (P), recall (R) and F-measures (F1) of the annotation aggregation with ROVER using LM scores from Leave-One-Utterance-Out (*LOUO*), Leave-One-Annotator-Out (*LOAO*), Leave-All-Annotators-Out (*LAAO*) settings. Baseline is a majority-voted ROVER. Additionally, we compute *Maximal* ROVER to compensate for low recall.

4. Conclusion

In this paper we have presented techniques for crowdsourcing complex annotation tasks: transfer of domain knowledge through priming and the annotation selection and aggregation techniques using joint stochastic language models. Transferring the domain knowledge and restricting the variability of annotator judgments through *priming* proved to be effective and increases the inter-annotator agreement and cross-language transfer performance of the workers. We have demonstrated that all the selection techniques are better than the random selection baseline. The utility of weighted aggregation of annotation hypotheses in the *targeted* crowdsourcing setting is low: majority-voting provides a very strong baseline. However, weighting each hypothesis with respect to the rest of crowdsourced annotations increases the F-measure by 0.4. Since the selection techniques we have proposed are unsupervised, they can be used as an online quality control mechanism.

5. Acknowledgements

This research is partially funded by the FP7 PortDial project n. 296170 and Spanish grants TIN2014-54288-C4-3-R and AP2010-4193.

6. References

[1] C. Callison-Burch and M. Dredze, "Creating speech and language data with Amazon's Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 1–12.

[2] M. Negri and Y. Mehdad, "Creating a bi-lingual entailment corpus through translations with Mechanical Turk: $100 for a 10-day rush," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 212–216.

[3] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 5270–5273.

[4] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2010, pp. 312–317.

[5] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1220–1229.

[6] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria," in *Proceedings of the NAACL HLT 2009 workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, 2009, pp. 27–35.

[7] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in Twitter data with crowdsourcing," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 80–88.

[8] G. Riccardi, A. Ghosh, S. A. Chowdhury, and A. O. Bayer, "Motivational Feedback in Crowdsourcing: A Case Study in Speech Transcription," in *Proceedings of the INTERSPEECH*, Lyon, France, August 2013, pp. 1111–1115.

[9] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[10] C. Qing, U. Endriss, R. Fernandez, and J. Kruger, "Empirical Analysis of Aggregation Methods for Collective Annotation," in *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, August 2014, pp. 1533–1542.

[11] J. Pustejovsky and A. Rumshisky, "Deep Semantic Annotation with Shallow Methods," LREC 2014 Tutorial, May 2014.

[12] S. Bangalore, G. Bordel, and G. Riccardi, "Computing Consensus Translation from Multiple Machine Translation Systems," in *In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001)*, 2001, pp. 351–354.

[13] Calvo, Marcos and García, Fernando and Hurtado, Lluís-F and Jiménez, Santiago and Sanchis, Emilio, "Exploiting multiple hypotheses for Multilingual Spoken Language Understanding," *CoNLL-2013*, pp. 193–201, 2013.

[14] E. A. Stepanov, I. Kashkarev, A. O. Bayer, G. Riccardi, and A. Ghosh, "Language Style and Domain Adaptation for Cross-Language SLU Porting," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Olomouc, Czech Republic: IEEE, December 2013, pp. 144–149.

[15] E. A. Stepanov, G. Riccardi, and A. O. Bayer, "The Development of the Multilingual LUNA Corpus for Spoken Language System Porting," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 2675–2678.

[16] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics," in *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece, 2009.

[17] S. A. Chowdhury, A. Ghosh, E. A. Stepanov, A. O. Bayer, G. Riccardi, and I. Klasinas, "Cross-Language Transfer of Semantic Annotation via Targeted Crowdsourcing," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, September 2014.