Cross-Language Transfer of Semantic Annotation via Targeted Crowdsourcing

Shammur Absar Chowdhury[1], *Arindam Ghosh*[1], *Evgeny A. Stepanov*[1],
Ali Orkan Bayer[1], *Giuseppe Riccardi*[1], *Ioannis Klasinas*[2]

[1]Dept. of Information Engineering & Computer Science, Univ. of Trento, Trento, Italy
[2]Dept. of Electronics & Computer Engineering, Tech. Univ. of Crete, Chania, Greece
{sachowdhury,aghosh,stepanov,bayer,riccardi}@disi.unitn.it, iklasinas@isc.tuc.gr

Abstract

The development of a natural language speech application requires the process of semantic annotation. Moreover multilingual porting of speech applications increases the cost and complexity of the annotation task. In this paper we address the problem of transferring the semantic annotation of the source language corpus to a low-resource target language via crowdsourcing. The current crowdsourcing approach faces several problems. First, the available crowdsourcing platforms have skewed distribution of language speakers. Second, speech applications require domain-specific knowledge. Third, the lack of reference target language annotation, makes crowdsourcing worker control very difficult. In this paper we address these issues on the task of cross-language transfer of domain-specific semantic annotation from an Italian spoken language corpus to Greek, via *targeted* crowdsourcing. The issue of domain knowledge transfer is addressed by priming the workers with the source language concepts. The lack of reference annotation is coped with a consensus-based annotation algorithm. The quality of annotation transfer is assessed using source language references and inter-annotator agreement. We demonstrate that the proposed computational methodology is viable and achieves acceptable annotation quality.

Index Terms: Crowdsourcing, Spoken language understanding, Annotation, Porting

1. Introduction

An important step in the development of a spoken language understanding system is semantic annotation of speech utterance transcriptions. A brute force approach to multilingual porting of speech applications would require the replication of this process for each target language. While the text of a corpus can be translated from the source language to the languages of interest using translation services, transfer of its annotation remains a research issue. Crowdsourcing – a recent computational model for large-scale distributed task execution – has the potential to be the solution. However, the feasibility of the semantic annotation via crowdsourcing is affected by factors such as the language of interest, the domain-specificity of the required annotation, and the availability of the resources for the evaluation of the crowd-annotated data.

First, the language of interest might be under-represented on existing crowdsourcing platforms due to the skewed worker demographics. Consequently, obtaining sufficient amount of adequately annotated data is an issue. An alternative is to access language speaker groups via other channels, and to design tasks *targeted* to that specific language groups.

Second, the semantic annotation required for speech applications is usually domain-specific. For example, for Information Technology domain a worker might be required to distinguish between hardware, software and network operations. Due to the fact that there is none to a minimal amount of time to provide some domain knowledge to the workers, the level of domain-specificity of the required annotation increases the complexity of the task, and it is expected to decrease the quality. Thus, the domain knowledge has to be transferred by other means. Researchers have successfully used live-feedback signals to improve the performance of the workers in crowdsourcing [1, 2]. In this paper, on the other hand, the workers are *primed* with the source language concepts.

Third, the traditional method for quality control in crowdsourcing tasks require the annotations to exist in a target language, which is not always the case. Coupled with the constraints imposed by the limited number of workers for low-resource languages, the traditional evaluation methodology is not applicable. We approach the problem of evaluation of annotation using inter-annotator agreement and the source language references. Traditional metrics for inter-annotator agreement are designed for a fixed number of annotator over a fixed dataset; thus, the evaluation of the quality of crowdsourced annotation without expert references is still an open question. Additionally, in cross-lingual tasks language distance is an important factor: since source language references may be reused for close languages and not for distant ones due to the word order and concept representation differences.

These issues are addressed on the task of cross-language transfer of domain-specific semantic annotation from Italian to Greek in spoken language corpus via *targeted* crowdsourcing. The language pair represents distant languages, which are under-represented on popular crowdsourcing platforms. The semantic annotation task requires workers to make two decisions – on the span of the concept and its label; thus, there is a task of *concept segmentation* as well as cross-language transfer of domain-specific *concept labels*.

The paper is structured as follows. In Section 2 we briefly review related works on cross-language annotation transfer and crowdsourced annotation. In Sections 3 and 4 we describe the DIY *targeted* crowdsourcing platform and the crowdsourced cross-language annotation transfer task design, respectively. Section 5 presents the evaluation methodology and the results. Section 6 provides concluding remarks.

2. Related Work

The cross-language annotation transfer in the literature was successfully applied to a variety of tasks via Statistical Machine Translation (SMT) methods [3, 4, 5, 6]. In the context of semantic annotation for spoken language application, the SMT

methodology was applied in [7] to transfer semantic annotation from French to Italian. The general idea of the approach is presented in Figure 1 that depicts Italian-Greek phrase alignment and the annotation transfer.

The annotation transfer via SMT requires parallel corpora, and its evaluation requires expert annotated resources. However, it is costly to obtain expert annotation for each language. To overcome this, we use crowdsourcing for cross-language transfer of semantic annotation and apply inter-annotator agreement as a measure of within target language annotation quality; and evaluate the annotations against source language references as a measure of cross-language transfer quality.

In recent years crowdsourcing has been successfully applied to a variety of research problems. The mechanism is usually ideal for performing tasks that can be broken into micro-tasks and distributed to a crowd of workers. In the Natural Language Processing (NLP) domain it has been used for corpus creation [8, 9] , transcription [10, 11], translation [12], and annotation tasks [13, 14]. On the other hand, we apply crowdsourcing for cross-language annotation transfer, which is different from general annotation, because the workers are provided with a set of concepts that exist in the utterance in the source language.

3. Targeted Crowdsourcing

The main challenge of generalistic human computation platforms such as Amazon Mechanical Turk is attracting a large number of qualified workers to participate in tasks while filtering out low quality workers and spammers. Since enrollment to such platforms does not require any particular skill set from workers, it is up to the task designers to overcome this issue. Traditionally, in research community this problem is solved using qualification tests, gold standard evaluation on selected items of the task [11], and other techniques to penalize low quality work. Additionally, the pseudo-anonymity of the workers enforced by most crowdsourcing platforms makes it difficult to target workers or worker-groups with the desired skill set.

Targeted crowdsourcing has evolved as a new paradigm with the intent to overcome this drawback. In targeted crowdsourcing the objective is to attract workers who are likely to have the skills needed for the target task and to design the platform appropriately. Crowdsourcing for creative ideas and problem solving are firm examples. For example, in enterprise settings, a crowd of employees was successfully used to improve the overall business process of the company [15]. Recently, the US Centers for Disease Control and Prevention (CDC) launched the CDCOLOGY project [16], a microvolunteering platform, targeting the population of registered university students. As an example of targeting a more special skill set, Open Mind Word Expert [17], a volunteer-based web framework to tag words with appropriate senses from WordNet, has been able to attract enough volunteers with sufficient proficiency for the tasks.

For the task of semantic annotation transfer from one language to another the required skill is the target language proficiency (Greek). The demographic distribution of workers on platforms such as Amazon Mechanical Turk is very skewed: close to 90% of turkers are from US and India [18]. Hence, the utility of the platform is low for NLP tasks involving languages of under-represented speaker groups. In collaboration with researchers from target language speaking institutions a targeted crowdsourcing experiment was carried out.

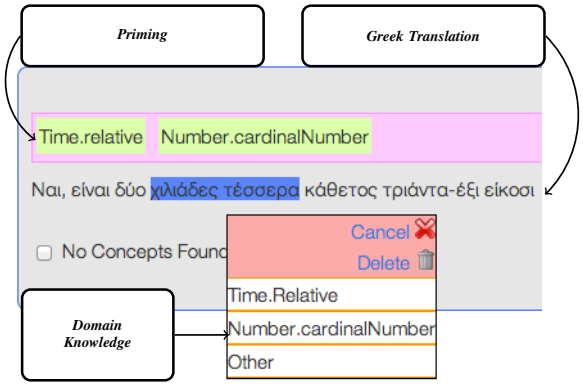

Figure 2: Description of each task. For each Greek utterance, the concepts from the source language (Italian) are used for priming. The domain knowledge is transferred using the LUNA concept ontology.

4. Semantic Annotation Transfer Task

The Multilingual LUNA Corpus [19, 20], was used for crowdsourced annotation transfer task. The corpus is the translation of Italian LUNA Corpus [21] to Spanish, Turkish and Greek via professional translation services. The translations are plain text, i.e. the semantic annotation have not been transferred.

4.1. Task Design

A set of 800 Greek utterances from the Multilingual LUNA Corpus [19] was put up for crowdsourcing. Each worker had to annotate 50 utterances presented on 5 pages (10 utterances per page).

The task had concise instructions and a short video demonstrating the annotation process to workers. Since Greek translations lack both segmentation and concept labels; the worker had to perform two subtasks: concept segmentation and labeling. After reading an utterance, a worker had to highlight a segment of an utterance covering a single concept and select the most suitable label from a drop-down menu (See Figure. 2).

The LUNA concept ontology contains a total of 45 unique concepts arranged in a two-level hierarchy with 26 top-level concepts. To ease the concept selection, the drop-down menu of concepts was arranged with respect to this 2-level hierarchy. No overlaps or nesting of concepts is allowed. However, a worker could mark an utterance as containing no concepts.

4.2. Priming the Workers

The semantic information is mostly preserved during the process of translation [5]. Consequently, the concepts from the Italian references were provided to the workers in the form of a unique list of suggested concepts on top of each utterance. The idea behind priming is to transfer the knowledge of the domain and provide a worker with semantic information to support the annotation task. The workers were free to highlight and mark segments matching the suggested concepts or ignore the list entirely.

5. Results and Discussion

In this Section we evaluate the quality of the annotations collected via crowdsourcing task described in the previous Section.

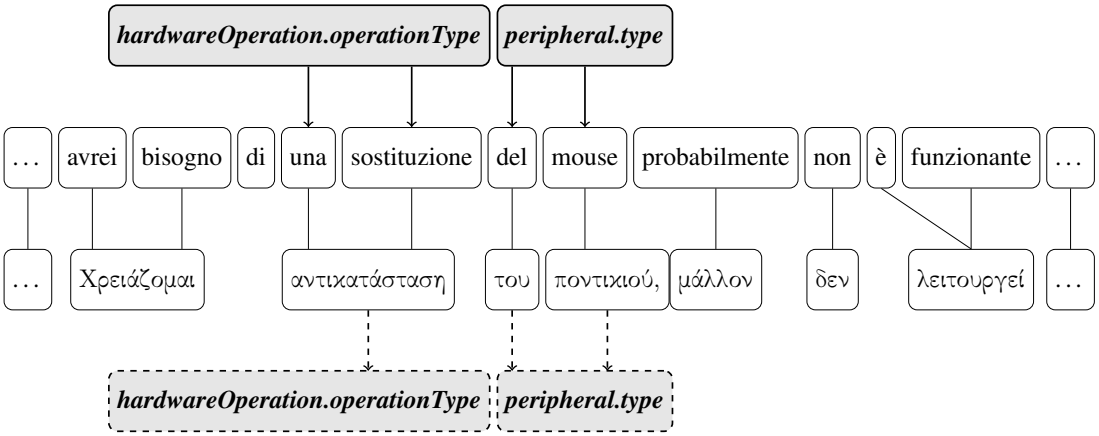

Figure 1: Cross-language annotation transfer task. Italian and Greek utterances are not one-to-one aligned. A concept can be linked to a single word in Greek, but multiple words in Italian or vice versa.

5.1. Data Collection Results

Fifty workers completed over 2000 micro-tasks over a period of two weeks. From the subset of 800 annotated utterances, 536 were annotated by at least three workers. The number of annotated concepts between languages differs: while there are 2,227 concepts in the references (Italian), there are on average 1,439 (35% less) concepts in Greek. Comparison between the suggested and the annotated concepts indicates that 44% of suggested concepts were ignored by the workers; while 9% of annotated concepts were not from the suggested lists.

For the evaluation we consider only utterances that have at least three judgments (536 utterances). We first evaluate the inter-annotator agreement between the workers, and then the transfer of annotation between languages.

5.2. Inter-Annotator Agreement

We first describe the evaluation methodology and then the agreement on the two subtasks of semantic annotation individually and together.

5.2.1. Evaluation Methodology

The commonly accepted metric for the assessment of the quality of an annotated resource is to measure the agreement between annotators. The most widely used agreement measure is κ (Cohen's for two and Fleiss' for several annotators), which is a chance corrected percent agreement measure. Unfortunately, κ is designed for a setting with a fixed number of annotators over a fixed data set; and this is not the case in crowdsourcing. Additionally, in text markup tasks, such as annotation, the number of *true negatives*, required for the calculation of the observed and chance agreements in κ, is not well defined (e.g. the number of text segments discarded by the workers as concept chunks). These factors make κ impractical as a measure of agreement of crowdsourced annotation.

An alternative agreement measure that does not depend on *true negatives* is Positive (Specific) Agreement [22], which is identical to the widely used F-measure [23]. Even though the measures are also for the fixed number of annotators on a common data set, since they do not rely on *true negatives* and the chance agreement, they are better suitable for the evaluation of crowdsourced annotation. In our crowdsourcing experiment we have collected 3 judgments per utterance; thus, for computing pair-wise F-measures we randomly assign each judgment to one of the three hypothetical annotators. The reported F-measures are averages of pair-wise F-measures among these three hypothetical annotators.

Match	**P**	**R**	**F1**
Whole Data			
Exact	39.60	38.58	39.08
Partial	64.24	62.90	63.56
Common Span Subset			
Exact	46.10	47.41	46.74
Partial	69.16	71.07	70.10

Table 1: Segmentation Agreement reported as averages of pair-wise precision (P), recall (R) and F-measures (F1) for exact and partial matches on whole data and the subset of common spans.

In text markup tasks annotators might select different spans all of which might be considered correct. For instance, for the *hardware* concept the selected span might be *with the printer*, *the printer*, or only *printer*. Thus, we report results for *exact* and *partial* matches [24]. Since in semantic annotation tasks workers are taking two decisions, we evaluate the agreement on these decisions separately as *segmentation* and *labeling* agreements and jointly as *semantic annotation agreement*.

5.2.2. Segmentation Agreement

Segmentation Agreement is the measure of the agreement of the workers on concept spans regardless of the label they give to the selected span. The averages of pair-wise precision, recall and F-measures are reported for exact and partially matched spans in Table 1 (upper part). Agreement on partial matches is relatively low: $F_1 = 63.56$, due to the fact that the measure also considers 'missing' concepts, i.e. identified only by one of the annotators. The segmentation agreement on the set of spans common to all of the judgments for an utterance is acceptably higher: $F_1 = 70.10$ (Table 1, lower part).

5.2.3. Labeling Agreement

Labeling Agreement is the measure of the agreement of the workers on the concept labels, regardless of the agreement on their spans. Unlike Segmentation Agreement there are no par-

	P	R	F1
Exact	48.39	47.15	47.76
Set	67.71	73.37	70.55

Table 2: Labeling Agreement reported as averages of pair-wise precision (P), recall (R) and F-measures (F1) for exact match and set (compares lists of unique concepts regardless of the order)

Match	P	R	F1
Exact	33.77	32.90	33.32
Partial	51.45	50.35	50.89

Table 3: Semantic Annotation Agreement – jointly for segmentation and labeling – reported as averages of pair-wise precision (P), recall (R) and F-measures (F1) for exact and partial matches.

tial matches (each concept is represented by a single token). In order to evaluate the labeling agreement independently from segmentation differences[1] we additionally compute the agreement over sets of annotated concepts.

The labeling agreement results are reported in Table 2. The average of pair-wise F-measures for the match (*Exact* in Table 2) is 47.76. The average of pair-wise F-measures for the set condition is considerably higher – 70.55. The results indicate that there are also differences in the segmentation of the same concepts.

5.2.4. Semantic Annotation Agreement

Semantic Annotation Agreement is the measure that considers both segmentation and labeling. It is the most strict of the inter-annotator agreement measures, since annotators have to agree both on the label and on its span. The results are reported in Table 3. The average of pair-wise F-measures for partial matches is only 50.89.

Even though, the inter-annotator agreement is relatively low on each of the subtasks of the semantic annotation, none of the workers is an expert. Thus, these results are indicative only of the variability in annotation. Since the task is a transfer of semantic annotation, there are also the expert annotated source language references. In the next Section we exploit these references to evaluate the quality of transfer and acceptability of the collected annotations.

5.3. Cross-Language Annotation Transfer

In this Section we evaluate the transfer of the annotation from the source language (Italian) to the target language (Greek). Similar to the previous subsection, we first present the evaluation methodology and then the results.

5.3.1. Evaluation Methodology

Since the order of concepts might be affected by the differences in the word-order between languages, the cross-language evaluation is carried on the sorted lists of concepts per utterance. We compare the annotated concept labels (i.e. spans are not considered) against the labels in the Italian reference preserving the number of concepts in each case. This evaluation allows us to assess the amount of actual transfer. For the evaluation we randomly select one of the judgments and compute precision, recall, and F-measure using Italian references. The procedure is repeated 1,000 times and the results are averaged.

Recognition Output Voting Error Reduction (ROVER) is one of the most frequently used tool in Automatic Speech Recognition community. The tool combines hypothesized sequence outputs of multiple recognition systems (in this case: workers) and selects the best scoring sequence. We applied the technique to the collected non-expert annotations to produce a single one. Since the three judgments are over the same utterance, we have applied majority voting on token level to decide on the span and the label of concepts (out-of-span tokens are taken as having 'null' label). As a result we obtain a single majority voted annotation hypothesis. Similar to random re-sampling, the output of ROVER is evaluated against Italian references. The expectation is that ROVER improves the overall annotation transfer.

[1] E.g.: a worker might choose to annotate numerical expressions like *one seven* as a single *number* concept or as two.

	P	R	F1
Random Re-sampling	84.40	54.54	66.26
ROVER	83.87	69.82	76.20

Table 4: Cross-Language Transfer using random re-sampling and ROVER as precision (P), recall (R) and F-measure (F1); for random re-sampling the results are averages of 1,000 iterations.

5.3.2. Quality of Transfer

The results for the two evaluation settings – random re-sampling and ROVER – are reported in Table 4. The results indicate that even with the inter-annotator agreement of $F_1 = 50.98$ for joint span and label decisions, using techniques such as ROVER, it is possible to exploit 'the power of the crowd' to transfer annotation with acceptable quality. By combining non-expert annotator decisions we gain approximately 15% in recall. Even though, the recall for transferred annotation using ROVER is ≈ 70, the precision is acceptably high ≈ 84.

Overall, the combination of crowdsourcing and computational techniques such as ROVER make the approach viable for the cross-language annotation transfer.

6. Conclusion

In this paper we have addressed the problem of transferring the semantic annotation from the source language corpus (Italian) to a low-resource distant target language (Greek) via crowdsourcing. We have addressed the issue of the skewed language speaker distribution of current crowdsourcing platforms by using targeted crowdsourcing. We have presented the approach to transfer domain knowledge, required for the semantic annotation, via priming with a list of source language concepts. Additionally, we have presented the methodology to assess quality of the crowd annotated corpora using inter-annotator agreement and evaluation against source language references. We have demonstrated that by combining the 'power of the crowd' in the form of multiple hypotheses with a computational method such as ROVER the resulting corpus achieves acceptable annotation quality.

7. Acknowledgments

This research has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 296170 – PortDial.

8. References

[1] S. Dow, A. Kulkarni, B. Bunge, T. Nguyen, S. Klemmer, and B. Hartmann, "Shepherding the crowd: managing and providing feedback to crowd workers," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011, pp. 1669–1674.

[2] G. Riccardi, A. Ghosh, S. Chowdhury, and A. O. Bayer, "Motivational feedback in crowdsourcing: A case study in speech transcription," in *Proceedings of Interspeech*, 2013.

[3] D. Yarowsky, G. Ngai, and R. Wicentowski, "Inducing multilingual text analysis tools via robust projection across aligned corpora," in *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001, pp. 1–8.

[4] E. Riloff, C. Schafer, and D. Yarowsky, "Inducing information extraction systems for new languages via cross-language projection," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 1–7.

[5] L. Bentivogli, P. Forner, and E. Pianta, "Evaluating cross-language annotation transfer in the multisemcor corpus," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 364.

[6] S. Padó and M. Lapata, "Cross-lingual annotation projection for semantic roles," *Journal of Artificial Intelligence Research*, vol. 36, no. 1, pp. 307–340, 2009.

[7] B. Jabaian, L. Besacier, and F. Lefèvre, "Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 3, 2013.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] M. Vukovic and A. Natarajan, "Operational excellence in it services using enterprise crowdsourcing," in *Services Computing (SCC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 494–501.

[16] CDCOLOGY, http://www.cdcology.sparked.com/, March 2014.

[17] T. Chklovski and R. Mihalcea, "Building a sense tagged corpus with open mind word expert," in *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*. Association for Computational Linguistics, 2002, pp. 116–122.

[18] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson, "Who are the turkers? worker demographics in amazon mechanical turk," *Department of Informatics, University of California, Irvine, USA, Tech. Rep*, 2009.

[19] 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 International Conference on Language Resources and Evaluation (LREC)*, 2014.

[20] E. A. Stepanov, I. Kashkarev, A. O. Bayer, G. Riccardi, and A. Ghosh, "Language style and domain adaptation for cross-language slu porting," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 144 – 149.

[21] 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 SRSL 2009 Workshop of EACL*, Athens, Greece, 2009.

[22] J. L. Fleiss, "Measuring agreement between two judges on the presence or absence of a trait," *Biometrics*, vol. 31, pp. 651–659, 1975.

[23] G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[24] R. Johansson and A. Moschitti, "Syntactic and semantic structure for opinion expression detection," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 2010, pp. 67–76.