Lang Resources & Evaluation
DOI 10.1007/s10579-017-9396-5

PROJECT NOTES

Cross-language transfer of semantic annotation via targeted crowdsourcing: task design and evaluation

Evgeny A. Stepanov[1] · Shammur Absar Chowdhury[1] · Ali Orkan Bayer[1] · Arindam Ghosh[1] · Ioannis Klasinas[2] · Marcos Calvo[3] · Emilio Sanchis[4] · Giuseppe Riccardi[1]

Abstract Modern data-driven spoken language systems (SLS) require manual semantic annotation for training spoken language understanding parsers. Multilingual porting of SLS demands significant manual effort and language resources, as this manual annotation has to be replicated. 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 cross-language semantic annotation transfer, may generate low judgment agreement and/or poor performance. The most serious issue in cross-language porting is the absence of reference annotations in the target

This research is partially funded by the EU FP7 PortDial Project No. 296170, FP7 SpeDial Project No. 611396, and Spanish contract TIN2014-54288-C4-3-R. The work presented in this paper was carried out while the author was affiliated with Universitat Politècnica de València.

✉ Evgeny A. Stepanov
evgeny.stepanov@unitn.it

Shammur Absar Chowdhury
shammar.chowdhury@unitn.it

Ali Orkan Bayer
aliorkan.bayer@unitn.it

Arindam Ghosh
arindam.ghosh@unitn.it

Ioannis Klasinas
iklasinas@isc.tuc.gr

Marcos Calvo
marcoscalvo@google.com

Emilio Sanchis
esanchis@dsic.upv.es

Giuseppe Riccardi
giuseppe.riccardi@unitn.it

Published online: 03 July 2017

language; thus, crowd quality control and the evaluation of the collected annotations is difficult. In this paper we investigate *targeted* crowdsourcing for semantic annotation transfer that delegates to crowds a complex task such as segmenting and labeling of concepts taken from a domain ontology; and evaluation using source language annotation. To test the applicability and effectiveness of the crowdsourced annotation transfer we have considered the case of close and distant language pairs: Italian–Spanish and Italian–Greek. The corpora annotated via crowdsourcing are evaluated against source and target language expert annotations. We demonstrate that the two evaluation references (source and target) highly correlate with each other; thus, drastically reduce the need for the target language reference annotations.

Keywords Crowdsourcing · Evaluation · Semantic annotation · Cross-language transfer

1 Introduction

With the increasing availability of intelligent digital assistants, spoken dialog systems (SDS) are at the forefront of research and development both in academia and industry. One of the main problems in the design of SDS for a multilingual user population and multi-domain applications is the cross-language porting process. Porting an existing SDS from one language to another essentially requires porting its language-specific components. In this paper we are interested in cross-language porting of spoken language understanding (SLU). The language understanding task requires, for each new target language, a mapping from word sequences to concept sequences or structures. This mapping has to take into account language differences while grounding speech transcriptions into a *shared* semantic representation of a task (e.g. travel reservation, open-domain personal assistant). We approach the problem using a crowdsourced semantic annotation transfer task. To test the applicability and effectiveness of the approach we consider the case of close and distant language pairs: Italian–Spanish and Italian–Greek.

Researchers and designers of spoken dialog systems have proposed semantic grammars to address the spoken language understanding (SLU) problem. Semantic grammars are formal models that *bind* the lexical representation and the concepts of a semantic representation. These models are usually based on hand-crafted rules and provide good performance for restricted tasks or dialogue contexts, e.g. (Rigo et al. 2009). More recently, the availability of very large speech and language corpora has

[1] Signals and Interactive Systems Lab, Department of Information Engineering and Computer Science, University of Trento, via Sommarive, 5, Trento, Italy

[2] Department of Electronics and Computer Engineering, Technical University of Crete, 731 00 Chania, Greece

[3] Google Switzerland, Brandschenkestrasse 110, Zurich 8002, Switzerland

[4] Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Camino de Vera s/n, 46020, Valencia, Spain

opened the opportunities for increased complexity and data-driven spoken language understanding, e.g. (Bayer and Riccardi 2012). Data-driven approaches have a superior performance and require less manual expertise, since they rely on availability of a corpus annotated with domain concepts. Porting a data-driven SLU involves generating annotated corpora in multiple languages while transferring the semantic representation of the source language task. In this paper we will consider the problem of cross-lingual porting of semantic annotations for a data-driven SLU task using crowdsourcing.

Recent advancement in large scale statistical machine translation (SMT) and the availability of off-the-shelf training tools have enabled the automation of the lexical cross-language porting. With respect to the direction and the object of translation, the approaches to spoken language understanding porting can be grouped under two categories: test-on-source and test-on-target. In the test-on-source approach the direction of translation is from a language the system is being ported to (target language) to the language of the existing SDS (source language); and the goal of the translation is to generate utterance transcriptions in the source language. Consequently, SLU of the existing system is 'extended' via SMT to cover a new language. The success of the approach depends on the quality of machine translation. In the test-on-target approach (also referred to as train-on-target), the direction of translation is the opposite, i.e. from the source language to the target language. In this case an SLU model is trained in target language based on the corpus generated by the source-to-target machine translation system and the semantic annotation transfer process.

In the literature, the test-on-source approach is credited as having better performance (Jabaian et al. 2010, 2011, 2013; Lefèvre et al. 2010; Calvo et al. 2016). The procedure is simpler to implement, since the only requirement is an SMT system. Moreover, Stepanov et al. (2013) have demonstrated that application of language-style and domain adaptation techniques to off-the-shelf and out-of-domain data trained SMT systems allows to improve their test-on-source SLU performance. Additional techniques such as statistical post-editing and 'smeared' SLU training proposed in (Jabaian et al. 2013); and re-ranking of the SLU hypotheses with in-domain joint language models trained on concept-word pairs proposed in (Stepanov et al. 2013) make this approach even more appealing. However, the test-on-target approach has its advantages, as it allows tuning and adaptation of the models in the target language directly, and it does not have an overhead of SMT during real-time execution.

The test-on-target approach relies on the automatic transfer of semantic annotation from the source to the target language. Starting from early 2000's, the annotation transfer (projection) approach was successfully applied to create monolingual annotated data for a variety of linguistic phenomena. Yarowsky et al. (2001) transferred annotations from English to close and distant languages and created resources for part-of-speech tagging (Xi and Hwa 2005), Noun phrase chunking, named-entity tagging and morphological analysis. Other applications include dependency parsing (Hwa et al. 2002), temporal annotation (Spreyer and Frank 2008), word sense disambiguation (Bentivogli et al. 2004), information extraction (Riloff et al. 2002), FrameNet (Padó and Lapata 2009), translation of

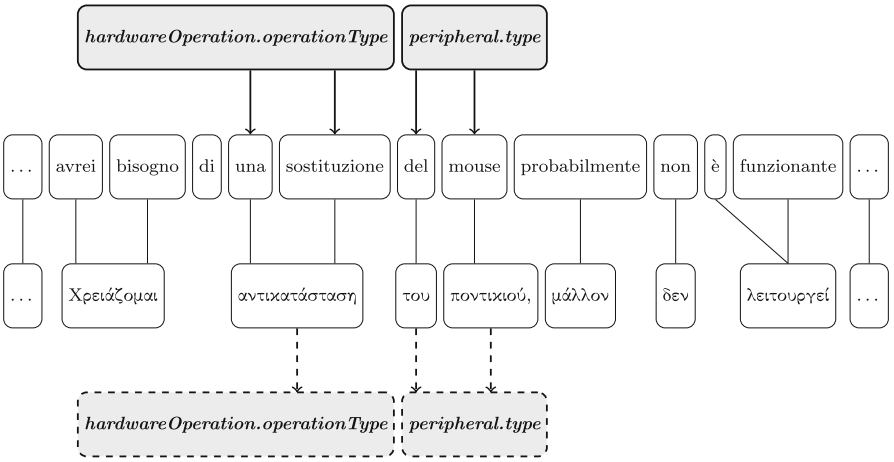

Fig. 1 General idea of cross-language annotation transfer using *indirect* alignment. Italian and Greek utterances are not one-to-one aligned. A concept can be linked to a single word in Greek, but multiple words in Italian or vice versa

Ontotext annotated biomedical patents (Gonzàlez et al. 2013), and others. In the context of spoken language understanding, the methodology was applied in (Jabaian et al. 2010, 2011, 2013) to transfer semantic annotation from French to Italian.

Jabaian et al. (2013) propose three annotation transfer approaches for spoken language understanding using statistical word alignments: (1) training alignments between source language concepts and target language utterances *directly*, (2) transferring source language annotation *indirectly* through word alignments, and (3) using SMT to translate text together with concept tags. The authors report *indirect* alignment having the best performance. However, due to the language differences, the lexical realization of concepts might differ across languages; and, as pointed out in (Jabaian et al. 2013), this reduces the applicability of the cross-language annotation transfer via statistical word alignments for distant language pairs (e.g. French–Arabic). With the rise of crowd computing—the combination of crowd intelligence and computational techniques—emerged an alternative to the annotation transfer via statistical word alignments. In this paper we propose a crowdsourcing task as a case of *direct* alignment for semantic annotation transfer. Since in crowdsourcing the annotation transfer is performed by humans, the approach avoids the issues of the SMT-based annotation transfer.

Complex tasks like semantic annotation transfer require workers to take simultaneous decisions on chunk segmentation and labeling, while acquiring domain-specific knowledge on-the-go. 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 by providing *semantic priming*. The general idea of the cross-language annotation transfer using *indirect* alignment is presented in Fig. 1 that depicts Italian–Greek phrase alignment and how concepts from the source language are mapped to the target language utterance. On the other hand, the

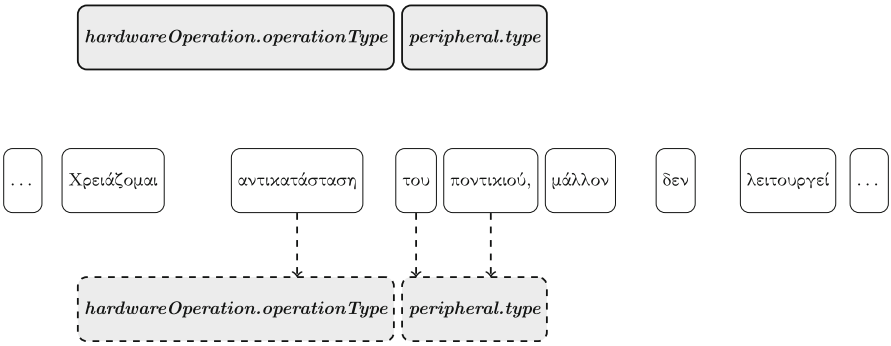

Fig. 2 Crowdsourced semantic annotation transfer with *priming* as a case of annotation transfer using *direct* alignment. The crowd aligns source language concepts and target language utterance tokens

proposed task of *primed* crowdsourced semantic annotation transfer through *direct* alignment is presented on Fig. 2. In the indirect alignment approach using crowdsourcing, the crowd needs to know both languages, which reduces the number of available workers. In the latter—direct alignment approach—the crowd generates alignments between source language concepts and the target language utterance tokens, without access to the original utterance. Consequently, the workers only need to know the target language; which allows to access a larger pool of workers.

Unfortunately, in the context of cross-language annotation transfer to low-resource languages, current crowdsourcing approaches face several limitations. Crowdsourcing platforms have a very skewed distribution of users; thus, the speakers of the desired low-resource language might be under-represented. Another limitation is that the lack of annotated target language references makes the quality control of workers difficult. We address these issues through the *targeted* crowdsourcing approach (Chowdhury et al. 2014), and evaluate the quality of the collected annotations using source language references and inter-annotator agreement. The adequacy of the source language references is evaluated as a correlation with the target language reference evaluation.

The paper is structured as follows. In Sect. 2 we describe the data set used for the annotation transfer task. The section also provides further description of the semantic annotation for spoken language understanding. In Sect. 3 we describe the concepts of *targeted* crowdsourcing. In Sect. 4 we describe the cross-language semantic annotation transfer methodology; and in Sect. 5 the targeted crowdsourcing task designed with respect to this methodology. In Sect. 6 we provide the methodology for inter-annotator agreement and cross-language annotation transfer evaluation. In Sect. 7 we describe the collected data and its evaluation. Section 8 provides concluding remarks.

2 Data set

The data set used throughout the paper for annotation transfer is the Multilingual LUNA Corpus (Stepanov et al. 2014), which is the professional translation of the human–machine dialogs of the Italian LUNA Corpus (Dinarelli et al. 2009) to Spanish, Turkish and Greek.[1]

The Italian LUNA Corpus (Dinarelli et al. 2009) is a collection of 723 human–machine (approximately 4000 turns and 5 h of speech) and 572 human–human (approximately 26,500 turns and 30 h of speech) spontaneous dialogs in the hardware/software help desk domain. The dialogs are conversations of the users involved in problem solving. While the human–human dialogs are recording of the real user-operator conversations, the human–machine dialogs are collected using the Wizard of Oz (WOZ) technique: the human agent (wizard) reacting to user requests is following one of the ten scenarios identified as most common by the help desk service provider. Text-to-speech synthesis (TTS) was used to provide responses to the users.

The attribute-value annotation of LUNA corpus uses a predefined ontology of concepts. There is an important distinction between the *attribute* of the concept, the *value* of the concept, and the lexical *span* of the concept.

Since the domain of the LUNA corpus is hardware/software help desk, the concepts are sets of domain-specific entities, such as *hardware*, *peripheral*, etc., and actions, such as *hardware operation*, *network operation*, etc.. The ontology also contains generic concepts such as *user*, *number*, *time*, etc.. The ontology consists of 45 unique concepts organized into two levels with the 26 top-level concepts. The second level of concepts can be seen as properties of the top-level concept. For example, for the top-level 'generic' concept *user*, the second level concepts are *name*, *surname*, *position*, *data*, etc.; for the top-level concept *computer*, the second level concepts are *type* (e.g. PC or laptop) and *brand* (e.g. DELL or HP). The two levels are usually considered together as an *attribute* of a concept. *Values* of concepts, on the other hand, in the *computer.type* example are *PC* or *laptop*. The *span* of the concept is the portion of an utterance string—a number of consecutive tokens—covered by the concept. The goal of this paper is to transfer the attribute-value annotation across languages using crowdsourcing.

The multilingual LUNA corpus consists of text only, i.e. annotations have not been transferred. While utterances in the corpus are in the source or target languages; the concept attribute annotation of the LUNA Corpus is in English and the ontology has not been translated.

Since the speaking style in the LUNA Corpus is conversational, the speech transcriptions include disfluencies such as repetitions, word repairs, etc. For the translations of the disfluencies, the professional translators were given two options. If the language pair is close enough to allow replicating disfluencies in the target language by the same morpho-syntactic means, without breaking the 'naturalness' of an utterance, they were replicated; and, if the speech disfluency in the target language requires a different morpho-syntactic operation (e.g. determiner or

[1] The corpora are available for research purposes from http://sisl.disi.unitn.it.

preposition repetition in the source language is translated as a content word, postposition or suffix repetition), the disfluency is marked in the text as such. As a result, speech disfluencies are replicated in Spanish, and are marked in Turkish and Greek. The multilingual LUNA corpus is intended as a reference resource for research on data-driven spoken language system porting; and it is used for the annotation transfer experiments in this paper.

3 Targeted crowdsourcing

Crowdsourcing is a task execution paradigm which endeavors to harness the knowledge and wisdom of the crowd to produce results comparable to that produced by domain experts. The most common form of crowdsourcing is the micro-tasking computational model, which involves dividing a large task into several small units, to be distributed to a crowd to perform. Researchers have successfully exploited online crowdsourcing platforms to show that non-expert crowds can match the performance of experts in natural language processing tasks such as speech transcription (Marge et al. 2010; Parent and Eskenazi 2010), translation (Zaidan and Callison-Burch 2011), and named entity annotation (Finin et al. 2010; Lawson et al. 2010) at a fraction of time and cost.

One major challenge while working with crowds on generalistic crowdsourcing platforms like Amazon Mechanical Turk[2] is attracting a large number of qualified workers to participate in tasks, while filtering out spammers and low quality workers. Since such online platforms usually do not have any in-place quality control for the workers, it is the responsibility of the task-designer to implement such checks. Most workers on such platforms lack skill-sets required to perform complex tasks, which might require domain knowledge. The pseudo-anonymity provided to the crowd-workers by these platforms makes it difficult for task-designers to target high-quality workers, or workers with a desired skill-set (Allahbakhsh et al. 2013; Fort et al. 2011). Traditionally, researchers have tried to solve this problem by designing quality controls such as qualification tests, gold standard evaluation on selected items of the task, and other techniques to penalize low quality work. A complex task like annotation transfer requires workers to exploit domain knowledge while taking simultaneous decisions on both segmentation and labeling. This increased task complexity may generate low judgment agreement and/or poor performance and may be unsuitable for in-the-wild crowdsourcing.

In targeted crowdsourcing, the objective is to attract crowd-workers who are likely to have the specialized skill sets and problem-solving expertise needed for the target task, and to design the platform appropriately. Usually such custom platforms are specifically designed for a particular task, targeted towards a particular group (or demographics) of users.

For the task of semantic annotation transfer from one language to another, the required skill is the target language proficiency (Spanish and Greek). The

[2] https://www.mturk.com.

demographic distribution of workers on platforms such as Amazon Mechanical Turk is very skewed: close to 90% of turkers are from US and India (Ross et al. 2009). Hence, the utility of the platform is low for NLP tasks involving languages of under-represented speaker groups. The targeted crowdsourced annotation task described in this paper was carried out in collaboration with the researchers from the target language speaking institutions, who advertised the annotation task to workers with the required language skills (i.e. proficiency in the target language and English). The cross-language semantic annotation transfer methodology and the design of the task used by the workers are described in the next sections.

4 Cross-language semantic annotation transfer methodology

As we defined in (Chowdhury et al. 2015), 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, that could be represented as a matrix H, such that:

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

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:

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

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.

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

where L is a finite set of task specific labels.

In case of a semantic annotation task, an utterance is annotated with a set of domain-specific concepts, such that a concept covers a certain span of an utterance; thus, the task consists of two sub-tasks: concept *segmentation* and *labeling*; and, essentially, 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 mappings is more complex.

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

The goal of a 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 support the training of machine learning algorithms. The most important factor for machine learning is consistency of the annotations. Thus, crowdsourced annotations must 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 (Pustejovsky and Rumshisky 2014). 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 (Pustejovsky and Rumshisky 2014), for instance. However, this simplification would require a lot more judgments to be collected. Thus, the best choice for the cross-language semantic annotation transfer 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 generally only the source language expert annotations are available, 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 cross-language annotation transfer 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. Additionally, the crowd can introduce new concepts from the ontology that are not present in the source language annotation. In the next section we present the targeted crowdsourcing task designed considering the proposed methodology.

5 Crowdsourced task design

Target language (Spanish and Greek) utterances from the Multilingual LUNA Corpus (Stepanov et al. 2014) were delivered for crowdsourcing. Each worker had to annotate 50 utterances presented on 5 pages (10 utterances per page).

The annotation task had concise instructions and a short video demonstrating the process to workers. Since translations lack both segmentation and concept labels, a worker had to perform two sub-tasks: concept segmentation and labeling. After reading an utterance, a worker had to highlight a segment of an utterance covering a

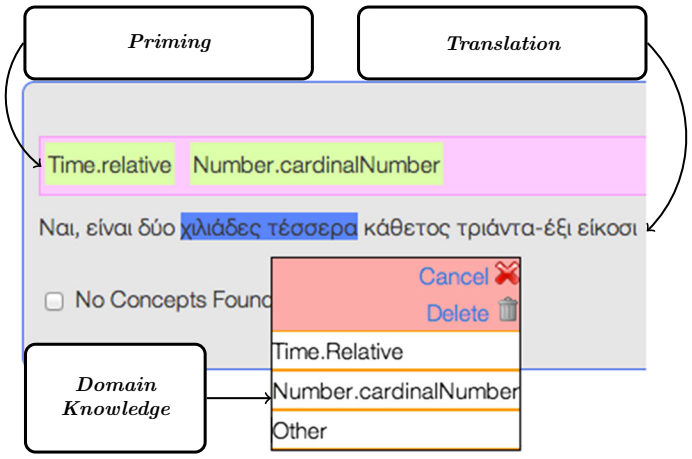

Fig. 3 Description of each task. For each target language utterance (Spanish or Greek), the concepts from the source language (Italian) are used for priming. The domain knowledge is transferred using the LUNA concept ontology

single concept and select the most suitable label from a drop-down menu (See Fig. 3).

As described in Sect. 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 is 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.

The domain knowledge transfer as *priming* with the concepts from the source language references is implemented in the form of a unique list of suggested concepts on top of each utterance. The list provides 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.

The expert target language annotations were collected using the same task setup. However, unlike the crowd, the experts had to annotate all the provided utterances.

6 Evaluation methodology

The task of cross-language transfer of semantic annotation via crowdsourcing requires two-way evaluation: consistency within and across languages. Within language consistency of the crowdsourced target language annotations is evaluated as inter-annotator agreement, whereas cross-language consistency is evaluated using standard information retrieval metrics of precision, recall and F-measure against the source language references.

In a realistic cross-language annotation transfer there are no target language references. In order to evaluate the adequacy of the source language references for the annotation transfer evaluation, we compare the crowdsourced annotations to

both the source and the target language references and measure correlation between the two.

6.1 Evaluation of inter-annotator agreement

The commonly accepted metric for the assessment of the quality of an annotated resource is the agreement between annotators. The most widely used agreement measure is κ—Cohen's (Cohen 1960) for two and Fleiss' (Fleiss 1971) 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 (P_o) and chance agreements (P_e) 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.

Equations 1–3 define Cohen's κ (Cohen 1960) and its observed (P_o) and chance (P_e) agreements in terms of *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN). In the equations N = TP + TN + FP + FN.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

$$P_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}} \tag{2}$$

$$P_e = \frac{\frac{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})}{\mathrm{N}} + \frac{(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}{\mathrm{N}}}{\mathrm{N}} \tag{3}$$

An alternative agreement measure that does not depend on *true negatives* is Positive (Specific) Agreement (Fleiss 1975) (P_{pos}, Eq. 4), also known as Dice's similarity coefficient (Dice 1945), which is identical to the widely used F_1-measure (Hripcsak and Rothschild 2005) (Eqs. 5–7). Even though Positive Agreement also requires a fixed number of annotators and a common data set, since it does not rely on *true negatives* and chance agreement, it is more suitable for the evaluation of a crowdsourced annotation (Chowdhury et al. 2014).

$$P_{pos} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

In the crowdsourcing experiments, we have collected 3 judgments per utterance; thus, for computing pair-wise F_1-measures we randomly assign each judgment to one of the three hypothetical annotators. The reported F_1-measures are averages of pair-wise F_1-measures among these three hypothetical annotators.

6.1.1 Exact and partial span matches

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 (Johansson and Moschitti 2010).

Partial matches are evaluated using 'soft' precision and recall metrics, as defined in (Johansson and Moschitti 2010). Unlike *exact* match evaluation, where *true positives* are counted only for spans that match the reference spans *exactly*; the 'soft' metrics consider the *coverage* of hypothesis spans. *Coverage* (c) of a span (s) is calculated with respect to another span (s'), as the number of tokens the two spans have in commons (intersection), as defined in Eq. 8, where |.| operator counts the number of tokens. If two spans have different labels, the coverage is set to zero

$$c(s, s') = \frac{|s \cap s'|}{|s'|}. \tag{8}$$

For the set of spans S, the authors define *span set coverage* C with respect to the set of spans S' according to Eq. 9

$$C(S, S') = \sum_{s_i \in S} \sum_{s_j \in S'} c\left(s_i, s'_j\right). \tag{9}$$

Precision and recall metrics are calculated with respect to the *span set coverage* according to Eqs. 10 and 11, where S_H and S_R are hypothesis and reference spans respectively, and |.| operator counts the number of spans.

$$\text{precision}(S_R, S_H) = \frac{C(S_R, S_H)}{|S_H|} \tag{10}$$

$$\text{recall}(S_R, S_H) = \frac{C(S_H, S_R)}{|S_R|} \tag{11}$$

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.

6.1.2 Segmentation agreement

Segmentation agreement is the measure of the agreement of the workers on concept spans regardless of the label they assign to the selected span. The averages of pair-wise precision, recall and F_1-measures are computed for *exact* and *partially* matched spans for all annotated concepts and a subset of concepts common to all annotators.

6.1.3 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 partial matches (each concept is represented by a single token). In order to evaluate the labeling agreement independently from segmentation differences (e.g.: a worker might choose to annotate numerical expressions like *one seven* as a single *number* concept or as two), we additionally compute the agreement over *sets* of annotated concepts (i.e. removing duplicates).

6.1.4 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. Similar to Segmentation Agreement, it is evaluated using pair-wise precision, recall and F_1-measures for *exact* and *partially* matched spans.

6.2 Evaluation of the quality of annotation transfer

The order of concepts in the source and the target languages might be affected by the differences in the word-order between languages. Moreover, segmentation of an utterance into concepts and their labeling might be affected by the languages' morphology and syntax. For example, semantic annotation transfer for a verbal *negation* concept from a language that expresses it as a word (e.g. English *not* or Italian *non*) to a language that expresses it as an affix (e.g. Turkish *-ma-*) is not possible without loss. Consequently, the accurate evaluation of the annotations generated via crowdsourcing requires target language references. Unfortunately, in a realistic annotation transfer scenario the target language references are not available.

An alternative is to use the source language references for the labeling evaluation. However, potential concept order differences due to the language distance need to be accounted for. Consequently, the cross-language evaluation is carried in different settings listed in Table 1.

For all the settings, we consider annotated concept labels (i.e. spans are not considered) against the labels in the source (Italian) references and the target language references. The two lists (hypothesis and reference) are aligned with

Table 1 Evaluation settings for the cross-language annotation transfer

	Description	Example
O	*O*riginal concept string	number number time number
C	Concept string after *C*onflation of the adjacent concepts with the same label (intended to account for segmentation differences between annotators and languages)	number time number
S	*S*orted concept string (intended to account for word order differences between languages)	number number number time
CS	Concept string after *C*onflation of the adjacent concepts with the same label and then *S*orting	number number time
SC	Concept string after *S*orting and *C*onflation of the adjacent concepts with the same label, i.e. *set* of concepts	number time

Operations of sorting (S) and conflation (C) of the adjacent concepts with the same label are applied to both hypothesis and reference concept strings

respect to Levenshtein distance and precision, recall, and F_1-measure are computed with respect to the alignment errors: insertions (I), deletions (D), and substitutions (S) according to the Eqs. 12–14.

$$\text{precision} = \frac{C}{C + I + S} \tag{12}$$

$$\text{recall} = \frac{C}{C + D + S} \tag{13}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{14}$$

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.

For the baseline evaluation—random re-sampling—we randomly select one of the collected judgments and compute precision, recall, and F_1-measure. The procedure is repeated 1000 times and the results are averaged.

As an alternative to random hypothesis selection, we also evaluate a single aggregate annotation hypothesis. Since the three judgments are over the same utterance, we are applying simple 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). All possible ties for majority voting are broken randomly; thus, similar to the random re-sampling, the procedure is repeated 1000 times and the results are averaged. Application of majority voting on the token level automatically conflates the hypotheses; consequently, it is compared to random re-sampling only for the settings with conflation (C, CS, and SC conditions described in Table 1). The expectation is that majority voting improves the overall annotation transfer.

6.3 Evaluation of the adequacy of the source language references

To evaluate the adequacy of the source language references for the evaluation of cross-language annotation transfer, we compute its correlation to the evaluation using the target language references using the bootstrap method. In the bootstrap method, we randomly select 300 judgments from all the judgments in crowdsourced data, without replacement. For the selected samples, we compute edit distance against the source and target language references. The procedure is repeated 10,000 times and Pearson's correlation coefficient (r) is computed between the two settings. To account for language distance, the correlation is computed for all the settings in Table 1 for both random re-sampling and majority voting. In case high positive correlation is observed between the evaluations against the two reference annotations (source and target), we will conclude that the source language references are adequate for the evaluation of the cross-language semantic annotation transfer.

7 Data analysis and evaluation results

In this section we first present the analysis of the collected data. Then, the evaluation of the data in terms of (1) the effect of priming as a method of constraining annotation variability, (2) inter-annotator agreement, and (3) annotation transfer, as described in Sect. 6. Last we present the evaluation of the adequacy of the source language references for the evaluation of annotation transfer.

7.1 Crowdsourced data analysis

For each target language, Spanish and Greek, 50 workers performed around 50 micro-tasks each for a period of 2 weeks. As described in Sect. 5, each micro-task consists of annotations comprising of concept segmentation and labeling for a single utterance. To build the consensus, we require at least 3 annotations per utterance. From the set of utterances provided for annotation, we obtained at least 3 annotations for 763 utterances in Spanish, and for 536 in Greek. From all the collected annotations, 416 utterances were annotated by both the crowd and the experts in both languages. To be able to compare the annotation results across languages, this common subset of 416 utterances was considered for evaluation (see Table 2).

Table 3 presents the analysis of the data collected through crowdsourced annotation on the common subset of the 416 utterances, 55 of which do not contain any concepts in the source language references. While there are 1579 concepts in the references (Italian), the reference annotation for Spanish ($\mathbf{ES}_e$) contains 1561 concepts and the reference annotation for Greek ($\mathbf{EL}_e$) contains 1091 concepts (31% less). For the crowdsourced annotation, on the other hand, on average there are 1070 (32% less) concepts in Spanish and 1027 (35% less) concepts in Greek.

The comparison between the suggested and the annotated concepts per judgment indicates that for Spanish 17% of suggested concepts were ignored by the expert, while 16% of the annotated concepts were not from the suggested lists; where as for Greek, 7% of suggested concepts were ignored by the expert, while 8% of annotated concepts were not from the suggested lists. The number of introduced new concepts is approximately the same for the experts and crowd workers; however, the number of ignored concepts is higher for crowd workers for both Spanish and Greek.

Analysis of the annotations reveals that for both languages—Spanish and Greek—the annotators tend to conflate annotations of consecutive *action.negation*[3] and *action* concepts as a single *action* or *action.negation* concept. Additionally, for Greek, the annotators tend to conflate consecutive *number* concepts into a single one. The observation explains the target language annotations' having less concepts than the source language annotation (the second row of Table 3). Moreover, specific to Spanish, the annotators prefer 'generic' concepts, like *problem-hardware* and *problem-general*, over specific ones these concepts can be decomposed to, such as *hardware-operation* and *hardware*. Consequently, for Spanish the number of

[3] Concept attribute names are simplified for readability.

Table 2 Amount of target language utterances annotated in Spanish (ES) and Greek (EL) by expert and crowd annotators

Data sets	Total
Expert annotation	746
Crowd annotation: ES	763
Crowd annotation: EL	536
Common subset	416
Non-primed annotation: ES	420

Common subset is a set of utterances annotated by the experts and the crowd in both languages

Table 3 Data annotation statistics for Spanish (**ES**) and Greek (**EL**) with respect to the Italian references (**IT**) for expert (*e*) and crowd (*c*) annotators in terms of annotated concepts in total and per worker judgment

	IT	ES_e	ES_c	EL_e	EL_c
# of concepts	1579	1561	1070	1091	1027
% of concepts w.r.t. source (IT)	100%	99%	68%	69%	65%
Suggested concepts list usage (per judgment)					
% of annotated concepts (w.r.t. Ref.)	100%	83%	74%	93%	82%
% of ignored concepts (w.r.t. Ref.)	0%	17%	26%	7%	18%
% of added concepts (w.r.t. Hyp.)	0%	16%	16%	8%	7%

For crowdsourced annotations ($\mathbf{ES}_c$ and $\mathbf{EL}_c$) values are averages across all judgments. While percent of annotated concepts is with respect to the total number of concepts, for suggested concept list usage percentages are with respect to the unique lists of concepts

ignored and added concepts with respect to the suggested list is higher than for Greek.

While these numbers are not indicative of the inter-annotator agreement, they are indicative of cross-language annotation transfer performance using the source language references: the lower number of annotated concepts in the target languages and the ratio of ignored concepts will impair recall, while the concepts that are not from the suggested list will impair precision. Since the percentage of both ignored and added concepts is higher for Spanish than for Greek, we expect better cross-language annotation transfer for the latter. As for the case of the evaluation using the target language references, we expect performance for Greek to be higher than for Spanish, since the percent of the used suggested concepts is higher for both the expert and the crowd annotators.

7.2 Crowdsourced data evaluation

For the analysis of the collected annotations, we first evaluate the effect of priming and then inter-annotator agreement between workers in the primed setting. The cross-language annotation transfer is evaluated last.

Table 4 Inter-annotator agreement for Spanish *primed* and *non-primed* annotation settings reported as averages of pair-wise precision (P), recall (R) and F_1-measures (F1) for the lists of unique concepts regardless of the order

	P	R	F1
Non-primed	0.369	0.341	0.354
Primed	0.622	0.560	0.590

Table 5 Cross-language transfer performance for Spanish *primed* and *non-primed* annotation settings using random re-sampling as averages of precision (P), recall (R) and F_1-measure (F1) of 1000 iterations

	P	R	F1
Non-primed	0.421	0.238	0.304
Primed	0.773	0.477	0.590

7.2.1 Semantic priming

As previously mentioned, the goal of priming in semantic annotation is two-fold: (1) to transfer the domain knowledge and (2) to constrain the word-to-concept mapping choices of the crowd. Thus, it is naturally expected that the annotation hypotheses collected in primed setting will have higher inter-annotator agreement, as well as be more consistent with the source language annotation.

An experiment comparing primed and non-primed settings is conducted for Spanish using 420 utterances from Multilingual LUNA Corpus (Stepanov et al. 2014). The inter-annotator agreements for both settings are given in Table 4 and the cross-language transfer performances using random re-sampling are given in Table 5. In both cases, the annotations collected using priming have much higher F_1-measures. Thus, we conclude that priming is effective for both domain knowledge transfer and restricting the mapping choices.

7.2.2 Inter-annotator agreement

In this section we provide results of the inter-annotator agreement evaluation—segmentation agreement, labeling agreement, and semantic annotation agreement.

Segmentation agreement measures the agreement between the workers on concept spans regardless of the label they give to the selected span. The averages of the pair-wise precision, recall and F_1-measures are reported for the exactly and partially matched spans for all concepts in Table 6 and for the matched concepts in Table 7. The agreement on the partial matches is 0.615 for Spanish and 0.654 for Greek, when all annotated concepts are considered, i.e. considering also 'missing' concepts identified only by one of the annotators. Whereas the segmentation agreement on the matched concept spans for all of the judgments for an utterance is

Table 6 Segmentation agreement reported as averages of pair-wise precision (P), recall (R) and F_1-measures (F1) for exact and partial matches on all concepts

	ES			EL		
	P	R	F1	P	R	F1
Exact	0.427	0.394	0.410	0.427	0.402	0.414
Partial	0.632	0.599	0.615	0.676	0.633	0.654

Table 7 Segmentation agreement reported as averages of pair-wise precision (P), recall (R) and F_1-measures (F1) for exact and partial matches on the set of matched concepts

	ES			EL		
	P	R	F1	P	R	F1
Exact	0.545	0.514	0.529	0.483	0.496	0.490
Partial	0.739	0.702	0.720	0.710	0.722	0.716

Table 8 Labeling agreement reported as averages of pair-wise precision (P), recall (R) and F_1-measures (F1) for exact match (O) and set (SC), that compares lists of unique concepts regardless of the order

	ES			EL		
	P	R	F1	P	R	F1
Exact						
(O)	0.500	0.461	0.480	0.523	0.494	0.508
Set						
(SC)	0.678	0.637	0.657	0.768	0.712	0.739

higher: 0.720 for Spanish and 0.716 for Greek. Overall, the segmentation agreement on the whole data and the set of matched concepts is similar across languages.

Labeling agreement measures the agreement of the workers on the concept labels, regardless of the agreement on their spans. The labeling agreement results are reported in Table 8. The average of pair-wise F_1-measures for the exact match (*Exact* in Table 8) is 0.480 for Spanish and 0.508 for Greek. The average of pair-wise F_1-measures for the set match condition is considerably higher—Spanish: 0.657 and Greek: 0.739. The results indicate that there are differences in the segmentation of the same concepts.

Semantic annotation agreement measures segmentation and labeling annotation jointly. 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 9 Semantic annotation agreement—jointly for segmentation and labeling—reported as averages of pair-wise precision (P), recall (R) and F_1-measures (F1) for exact and partial matches

	ES			EL		
	P	R	F1	P	R	F1
Exact	0.370	0.341	0.355	0.367	0.346	0.357
Partial	0.515	0.482	0.498	0.555	0.520	0.537

Table 9. The average of pair-wise F_1-measures for partial matches is only 0.498 for Spanish (ES), and 0.537 for Greek (EL).

The inter-annotator agreement for each of the sub-tasks of the semantic annotation indicates the variability in annotation between the non-expert annotators, which also indicates the complexity of the semantic annotation transfer task. Since the task is to transfer the semantic annotation of the source language to a target language, we have the expert annotated source and target language references; thus, next we exploit these references to evaluate the quality of transfer and acceptability of the collected annotations.

7.2.3 Evaluation of annotation transfer

The availability of the target language references allows us to estimate the upper bound of the annotation transfer performance via crowdsourcing. To estimate the upper bound, we compute *labeling agreement* as precision, recall and F_1 between the source and the target language references annotations. As previously mentioned, for Spanish expert annotations, the number of ignored and added concepts are higher than for Greek, despite the fact that for Greek less concepts were annotated. Thus, we expect the Greek expert annotations to have higher agreement with Italian than the Spanish annotations.

The labeling agreement for the reference annotations is reported in Table 10 (*Expert agr.* row) for each of the evaluation settings defined in Table 1. Overall, the agreement between the source Italian annotations and the expert target language annotations is good. For both languages the best agreement is observed for the *SC* setting (i.e. *set*): $F_1 = 0.777$ for Spanish and $F_1 = 0.926$ for Greek. The difference between the two languages is evident in the fact that for Greek the agreements are higher for the *conflated* evaluation settings (C and CS), whereas for Spanish they are higher for the settings without *conflation*, i.e. the original (O) and the sorted (S) concept strings. As expected, for Greek the agreement is higher for all the evaluation settings.

The results for the two evaluation settings for crowdsourced annotation—random re-sampling and majority voting—against the source and the target language references are reported in Table 10. For majority voting, we report only conflated results (i.e. C, CS, and SC), as the technique conflates the adjacent concepts with the same label into a single one.

Table 10 Annotation transfer as F_1-measure for random re-sampling (*RandRS*) and majority voting (*MV*) evaluated against the source language (**SRC**) Italian references and the target language (**TGT**) Spanish (**ES**) or Greek (**EL**) references under the evaluation settings reported in Table 1: original concept string (O), and using operations of Sorting (S) and Conflation (C) of the adjacent concepts with the same label

	ES					EL				
	O	C	S	CS	SC	O	C	S	CS	SC
Expert agr.										
	0.729	0.688	0.750	0.725	0.777	0.767	0.859	0.787	0.879	0.926
SRC										
RandRS	0.614	0.638	0.645	0.675	0.728	0.661	0.729	0.690	0.762	0.820
MV	–	0.674	–	0.716	0.777	–	0.770	–	0.798	0.861
TGT										
RandRS	0.617	0.625	0.627	0.638	0.683	0.679	0.717	0.707	0.750	0.784
MV	–	0.657	–	0.669	0.718	–	0.753	–	0.773	0.816

The *Expert arg.* row reports the expert annotator agreement for the same evaluation settings, and represents the upper-bound of the crowdsourced semantic annotation transfer

The first observation is that performances of the majority voting output are higher than the random re-sampling for both Spanish and Greek. The observation indicates that the combination of crowdsourcing with computational techniques is useful for the cross-language annotation transfer. The second observation is that the crowd performance is below the expert agreement with the source language reference annotation for both languages, except the SC (*set*) setting for Spanish, where the crowd performance reaches the upper-bound of expert agreement ($F_1 = 0.777$). For Greek, generally, all the performances are higher. The difference is predicted from Table 3, as Greek crowdsourced data has less ignored and added concepts, as well as the numbers are closer to that of the expert annotations. The third observation is that the performance differences are preserved in the evaluation against the source and the target language references. Thus, the source language references alone are sufficient for the estimation of the crowd performance. In the next section we evaluate the correlation between the evaluations using the source and the target language references.

7.3 Correlation of the source and the target semantic reference annotations

As previously mentioned, we use the bootstrap method to randomly select 300 judgments from all the judgments in crowdsourced data, without replacement, and repeat the procedure 10,000 times. The correlation performance is reported in Table 11. The values in the table are indicative of several factors.

Similar to the annotation transfer performance, for Spanish, the highest correlation is for the original concept string and the sorted concept string. Moreover, the values are close to each other: 0.73 and 0.75. The high correlation for these two settings indicates the word order closeness of the two languages, as well

Table 11 Correlation of performance between using expert source and target language references as Pearson's *r* for 300 random utterances for random re-sampling (*RandRS*) and majority voting (*MV*)

	ES					EL				
	O	C	S	CS	SC	O	C	S	CS	SC
RandRS	0.73	0.64	0.75	0.66	0.66	0.42	0.79	0.39	0.76	0.80
MV	–	0.66	–	0.67	0.63	–	0.80	–	0.79	0.79

Scores are averages of 10,000 iterations

as similarity of Spanish and Italian reference annotations. For Greek, on the other hand, the highest correlation is observed for the settings that apply conflation (C, CS, and SC). This correlation is also predictable from the concept count differences between the Italian and Greek references.

The correlation for majority voting is higher than the correlation for random re-sampling for both languages for all the evaluation settings except SC. The difference for Greek is negligible: 0.80 versus 0.79, for random re-sampling and majority voting, respectively. For Spanish the difference is higher: 0.66 versus 0.63. Table 11 suggests that it is better to conflate the hypotheses for Greek (C, CS, or SC), but not for Spanish. As for the original and the sorted concepts string settings, for Greek there are weak positive correlations (0.42 and 0.39 for random re-sampling), and for Spanish the correlations are high (0.73 and 0.75). Overall, the high positive correlation between evaluations using the source and the target language references for the selected settings per language supports the adequacy of the cross-language transfer evaluation using the source language references.

8 Conclusion

In this paper we have addressed the problem of transferring the semantic annotation from the source language corpus (Italian) to close and distant languages—Spanish and 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 domain and language independent approach to transfer domain knowledge, required for the semantic annotation, via priming with the source language concepts. Additionally, we have presented the methodology to assess the quality of the crowd annotated corpora using inter-annotator agreement and evaluation against the source language references. We have demonstrated that by combining the 'power of the crowd' in the form of multiple hypotheses with a computational methods the resulting corpus achieves acceptable annotation quality. Most importantly, we have evaluated the adequacy of the source language references for the evaluation of cross-language annotation transfer, and found a high correlation between the source and the target language reference evaluation. Thus, we have demonstrated that for the evaluation of cross-language porting it is sufficient to have the source language references only, avoiding the effort of collecting the target language reference annotations.

References

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems. *IEEE Internet Computing*, *17*(2), 76–81.

Bayer, A. O., & Riccardi, G. (2012). Joint language models for automatic speech recognition and understanding. In *Proceeding of the IEEE spoken language technology workshop*.

Bentivogli, L., Forner, P., & Pianta, E. (2004). Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the 20th international conference on computational linguistics, association for computational linguistics*.

Calvo, M., Hurtado, L. F., Garcia, F., Sanchis, E., & Segarra, E. (2016). Multilingual spoken language understanding using graphs and multiple translations. *Computer Speech and Language*, *38*, 86–103.

Chowdhury, S. A., Calvo, M., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., et al. (2015). Selection and aggregation techniques for crowdsourced semantic annotation task. In *The 16th annual conference of the international speech communication association (INTERSPEECH)* (pp. 2779–2783). Dresden: ISCA.

Chowdhury, S. A., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., & Klasinas, I. (2014). Cross-language transfer of semantic annotation via targeted crowdsourcing. In *The 15th annual conference of the international speech communication association (INTERSPEECH)* (pp. 2108–2112). Singapore: ISCA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Dinarelli, M., Quarteroni, S., Tonelli, S., Moschitti, A., & Riccardi, G. (2009). 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.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). 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* (pp. 80–88).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651–659.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–420.

Gonzàlez, M., Mateva, M., Enache, R., Na, C. E., Arquez, L. M., Popov, B., & Ranta, A. (2013). MT techniques in a retrieval system of semantically enriched patents. In MT Summit.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298.

Hwa, R., Resnik, P., Weinberg, A., & Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th annual meeting on association for computational linguistics, association for computational linguistics* (pp. 392–399).

Jabaian, B., Besacier, L., & Lefèvre, F. (2010). Investigating multiple approaches for SLU portability to a new language. In *Proceedings of INTERSPEECH*.

Jabaian, B., Besacier, L., & Lefèvre, F. (2011). Combination of stochastic understanding and machine translation systems for language portability of dialogue systems. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)*.

Jabaian, B., Besacier, L., & Lefèvre, F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(3), 636–648.

Johansson, R., & Moschitti, A. (2010). Syntactic and semantic structure for opinion expression detection. In *Proceedings of the 40th conference on computational natural language learning* (pp. 67–76).

Lawson, N., Eustice, K., Perkowitz, M., & Yetisgen-Yildiz, M. (2010). Annotating large email datasets for named entity recognition with 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* (pp. 71–79).

Lefèvre, F., Mairesse, F., & Young, S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *Proceedings of INTERSPEECH*.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the amazon mechanical turk for transcription of spoken language. In *2010 IEEE international conference on acoustics speech and signal processing (ICASSP)* (pp. 5270–5273). IEEE.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, *36*(1), 307–340.

Parent, G., & Eskenazi, M. (2010). Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *IEEE spoken language technology workshop (SLT)* (pp. 312–317). IEEE.

Pustejovsky, J., & Rumshisky, A. (2014). *Deep semantic annotation with shallow methods*. LREC 2014 Tutorial.

Rigo, S., Stepanov, E. A., Roberti, P., Quarteroni, S., & Riccardi, G. (2009). The 2009 UNITN EVALITA Italian spoken dialogue system. In *Evaluation of NLP and speech tools for Italian workshop (EVALITA)*. Reggio Emilia.

Riloff, E., Schafer, C., & Yarowsky, D. (2002). 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* (pp. 1–7).

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). *Who are the turkers? Worker demographics in amazon mechanical turk*. Tech Rep. Irvine: Department of Informatics: University of California.

Spreyer, K., & Frank, A. (2008). Projection-based acquisition of a temporal labeller. In *Proceedings of the international joint conference on natural language processing* (pp. 489–496).

Stepanov, E. A., Kashkarev, I., Bayer, A. O., Riccardi, G., & Ghosh, A. (2013). Language style and domain adaptation for cross-language SLU porting. In *IEEE workshop on automatic speech recognition and understanding (ASRU)* (pp. 144–149). Olomouc: IEEE.

Stepanov, E. A., Riccardi, G., & Bayer, A. O. (2014). The development of the multilingual LUNA corpus for spoken language system porting. In *The 9th international conference on language resources and evaluation (LREC'14)* (pp. 2675–2678). Reykjavik: ELRA.

Xi, C., & Hwa, R. (2005). A Backoff model for bootstrapping resources for non-english languages. In *Proceedings of the conference on human language technology and empirical methods in natural language processing, association for computational linguistics* (pp. 851–858).

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st international conference on human language technology research, association for computational linguistics* (pp. 1–8).

Zaidan, O. F., & Callison-Burch, C. (2011). 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* (pp. 1220–1229).