Creating and Validating Multilingual Semantic Representations for Six Languages: Expert versus Non-Expert Crowds

Mahmoud El-Haj, Paul Rayson, Scott Piao and Stephen Wattam
School of Computing and Communications, Lancaster University, Lancaster, UK
initial.surname@lancaster.ac.uk

Abstract

Creating high-quality wide-coverage multilingual semantic lexicons to support knowledge-based approaches is a challenging time-consuming manual task. This has traditionally been performed by linguistic experts: a slow and expensive process. We present an experiment in which we adapt and evaluate crowdsourcing methods employing native speakers to generate a list of coarse-grained senses under a common multilingual semantic taxonomy for sets of words in six languages. 451 non-experts (including 427 Mechanical Turk workers) and 15 expert participants semantically annotated 250 words manually for Arabic, Chinese, English, Italian, Portuguese and Urdu lexicons. In order to avoid erroneous (spam) crowdsourced results, we used a novel task-specific two-phase filtering process where users were asked to identify synonyms in the target language, and remove erroneous senses.

1 Introduction

Machine understanding of the meaning of words, phrases, sentences and documents has challenged computational linguists since the 1950s, and much progress has been made at multiple levels. Different types of semantic annotation have been developed, such as word sense disambiguation, semantic role labelling, named entity recognition, sentiment analysis and content analysis. Common to all of these tasks, in the supervised setting, is the requirement for a wide coverage semantic lexicon acting as a knowledge base from which to select or derive potential word or phrase level sense annotations.

The creation of large-scale semantic lexical resources is a time-consuming and difficult task. For new languages, regional varieties, dialects, or domains the task will need to be repeated and then revised over time as word meanings evolve. In this paper, we report on work in which we adapt crowdsourcing techniques to speed up the creation of new semantic lexical resources. We evaluate how efficient the approach is, and how robust the semantic representation is across six languages.

The task that we focus on here is a particularly challenging one. Given a word, each annotator must decide on its meaning[s] and assign the word to single or multiple tags in a pre-existing semantic taxonomy. This task is similar to that undertaken by trained lexicographers during the process of writing or updating dictionary entries. Even for experts, this is a complex task. Kilgarriff (1997) highlighted a number of issues related to lexicographers 'lumping' or 'splitting' senses of a word and cautioned that even lexicographers do not believe in words having a "discrete, non-overlapping set of senses". Véronis (2001) showed that inter-annotator agreement is very low in sense tagging using a traditional dictionary. For our purpose, we use the USAS taxonomy.[1] If a linguist were undertaking this task, as they have done in the past with Finnish (Löfberg et al., 2005) and Russian (Mudraya et al., 2006) USAS taxonomies, they would first spend some time learning the semantic taxonomy. In this experimental scenario, we aim to investigate whether or not non-expert native speakers can succeed on the word-to-senses classification task without being trained on the taxonomy in advance, therefore mitigating a significant overhead for the work. In addition, further motivation for our experiments is to validate the applicability of the USAS taxonomy (Rayson et

[1] The UCREL Semantic Analysis System (USAS), see http://ucrel.lancaster.ac.uk/usas/

Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, pages 61–71,
Valencia, Spain, April 4 2017. ©2017 Association for Computational Linguistics

al., 2004), with a non-expert crowd, as a framework for multilingual sense representation. The USAS taxonomy was selected for this experiment since it offers a manageable coarse-grained set of categories that have already been applied to a number of languages. This taxonomy is distinct from other potential choices, such as WordNet. The USAS tagset is originally loosely based on the Longman Lexicon of Contemporary English (McArthur, 1981) and has a hierarchical structure with 21 major domains (see table 1) subdividing into three levels. Versions of the USAS tagger or tagset exist in 15 languages in total and for each language, native speakers have re-evaluated the applicability of the tagset with some specific extensions for Chinese (Qian and Piao, 2009) but otherwise the tagset is stable across all languages. For each language tagger, separate linguistic resources (lexicons) have been created, but they all use the same taxonomy.

Domain	Description
A	General and abstract terms
B	The body and the individual
C	Arts and crafts
E	Emotion
F	Food and farming
G	Government and public
H	Architecture, housing and the home
I	Money and commerce in industry
K	Entertainment, sports and games
L	Life and living things
M	Movement, location, travel and transport
N	Numbers and measurement
O	Substances, materials, objects and equipment
P	Education
Q	Language and communication
S	Social actions, states and processes
T	Time
W	World and environment
X	Psychological actions, states and processes
Y	Science and technology
Z	Names and grammar

Table 1: USAS top level semantic fields

In terms of main contributions, our research goes beyond the previous work on crowdsourcing word meanings which requires workers to pick a word sense from an existing list that matches provided contextual examples, such as a concordance list. In our work, we require the participants to define the list of all possible senses that a word could take in different contexts. We also see that our two-stage filtering process tailored for this task helps to improve results. We compare interrater scores for two groups of experts and non-experts to examine the feasibility of extracting high-quality semantic lexicons via the untrained crowd. Non-experts achieved results between 45-97% for accuracy, between 48-92% for completeness, with an average of 18% of tasks having erroneous senses being left in. Experts scored 64-96% for accuracy, 72-95% for completeness, but achieve better results in terms of only 1% of erroneous senses left behind. Our experimental results show that the non-expert crowdsourced annotation process is of a good quality and comparable to that of expert linguists in some cases, although there are variations across different languages. Crowdsourcing provides a promising approach for the speedy generation and expansion of semantic lexicons on a large scale. It also allows us to validate the semantic representations embedded in our taxonomy in the multilingual setting.

2 Related Work

The crowdsourcing approach, in particular Mechanical Turk (MTurk), has been successfully applied for a number of different Natural Language Processing (NLP) tasks. Alonso and Mizzaro (2009) adopted MTurk for five types of NLP tasks, resulting in high agreement between expert gold standard labels and non-expert annotations, where a small number of workers can emulate an expert. With the possibility of achieving good results quickly and cheaply, MTurk has been tested for a variety of tasks, such as image annotation (Sorokin and Forsyth, 2008), Wikipedia article quality assessment (Kittur et al., 2008), machine translation (Callison-Burch, 2009), extracting key phrases from documents (Yang et al., 2009), and summarization (El-Haj et al., 2010). Practical issues such as payment and task design play an important part in ensuring the quality of the resulting work. Many designers pay between $0.01 to $0.10 for a task taking a few minutes. Quality control and evaluation are usually achieved through confidence scores and gold-standards (Donmez et al., 2009; Bhardwaj et al., 2010). Past research has shown (Aker et al., 2012) that the use of radio button design seems to lead to better results compared to the free text design. Particularly important in our case is the language demographics of MTurk (Pavlick et al., 2014), since we need to find enough native speakers in a number of languages.

There is a growing body of crowdsourcing work related to semantic annotation. Snow et al. (2008) applied MTurk to the Word Sense Disambiguation (WSD) task and achieved 100% precision with simple majority voting for the correct sense of the word 'president' in 177 example sentences. Rumshisky et al. (2012) derived a sense inventory and sense-annotated corpus from MTurkers comparison of senses in pairs of example sentences. They used clustering methods to identify the strength of coders' tags, something that is poorly suited to rejecting work from spammers (participants who try to cheat the system with scripts or random answers) and would likely not transfer well to our experiment.

Akkaya et al. (2010) also performed WSD using MTurk workers. They discuss a number of methods for ensuring quality, accountability, and consistency using 9 tasks per word and simple majority voting. Kapelner et al. (2012) increased the scale to 1,000 words for the WSD task and found that workers repeating the task do not learn without feedback. A set-based agreement metric was used by Passonneau et al. (2006) to assess the validity of polysemous selections of word senses from WordNet categories. Their objective was to take into account similarity between items within a set, however, this may not be desirable in our case due to the limited depth of the USAS taxonomy.

Directly related to our research here are the experiments reported in Piao et al. (2015). A set of prototype semantic lexicons were automatically generated by transferring semantic tags from the existing USAS English semantic lexicon entries to their translation equivalents in Italian, Chinese and Portuguese via dictionaries and bilingual lexicons. While some dictionaries involved, including Chinese/English and Portuguese/English dictionaries, provided high quality lexical translations for core vocabularies of these languages, the bilingual lexicons, including FreeLang English/Italian, English/Portuguese lexicons[2] and LDC English/Chinese word list, contain erroneous and inaccurate translations. To reduce the error rate, some manual cleaning was carried out, particularly on the English-Italian bilingual lexicons. Because of the substantial amount of time needed for such manual work, the rest of the lexical resources were used with only minor sporadic manual checking. Due to the noise introduced from the bilingual lexicons, as well as the ambiguous nature of the translation, the automatically generated semantic lexicons for the three languages contain errors, including erroneous semantic tags caused by incorrect translations, and inaccurate semantic tags caused by ambiguous translations. When these automatically generated lexicons were integrated and applied in the USAS semantic tagger, the tagger suffered from error rates of 23.51%, 12.31% and 10.28% for Italian, Chinese and Portuguese respectively.

The improvement of the semantic lexicons is therefore an urgent and challenging task, and we hypothesise that the crowdsourcing approach can potentially provide an effective means for addressing this issue on a large scale, while at the same time allowing us to further validate the representation of word senses in the USAS sense inventory (i.e. the semantic tagset) for these languages.

3 Semantic Labeling Experiment

We test the wisdom of the crowd in building lexicons and applying the same multilingual semantic representation in six languages: Arabic, Chinese, English, Italian, Portuguese and Urdu. These languages were selected to provide a range of language families, inflectional and derivational morphology, while covering significant number of speakers worldwide. For each language, we randomly selected 250 words. All experiments presented here use the USAS taxonomy to describe semantic categories[3] (Rayson et al., 2004).

3.1 Gold Standard Semantic Tags

To prepare gold standard data we asked a group of linguists (up to two per language) to manually check 250 randomly selected words for each of the six languages, starting from the data provided by Piao et al. (2015). For additional languages (those not in Piao et al. (2015)), the Arabic and Urdu gold standards were completed manually by native speaker linguists who translated the 250 words, and we instructed the translators to opt for the most familiar Arabic or Urdu equivalents

[2] http://www.freelang.net/dictionary/

[3] http://ucrel.lancaster.ac.uk/usas/semtags.txt

of the English words. This was further confirmed by checking the list of words by two other Arabic and Urdu native speakers. Here the base form of verbs in Arabic is taken to be the present simple form of the verb in the interest of convenience and because there is no part of speech tag for ‘present tense of a lexical verb’. Hence, the three-letter past tense verbs are tagged as ‘past form of lexical verb’, rather than as base forms. Also, while present and past participles (e.g., ‘interesting’, ‘interested’) are tagged as adjectives in English, these are labeled in Arabic as nouns in stand-alone positions but they can also function as adjective premodifying nouns. Linguists then used the USAS semantic tagset to semantically label each word with the most suitable senses.

3.2 Non-expert Participants

Non-expert participants are defined as those who are not familiar with the USAS taxonomy in advance of the experiment. We selected Amazon’s Mechanical Turk[4] – an online marketplace for work that requires human intelligence – and published “Human Intelligence Tasks” (HITs) for English, Chinese and Portuguese only. For Arabic, Italian, and Urdu initial experiments using MTurk showed that not enough native speakers are available to complete the tasks. Therefore, we employed 12 non-expert participants directly with four native speakers for each of the three languages. All participants used the same interface (Figure 1).

On MTurk, we paid workers an average of 7 US dollars per hour. We paid Portuguese workers 50% more to try and attract more participants due to the lack of Portuguese native speakers on MTurk. We paid the other directly contacted participants an average of 8 British pounds per hour. Those payments were made using Amazon[5] and Apple iTunes[6] vouchers.

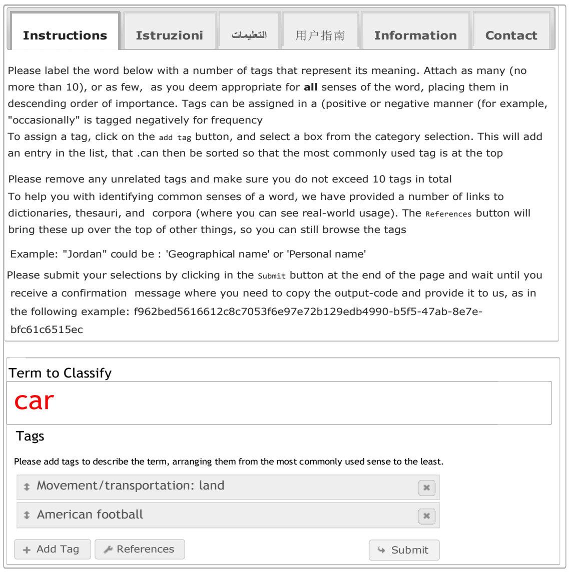
3.3 Expert Participants

Expert participants are defined as those who were already familiar with the USAS taxonomy before the experiments took place. For four languages (Arabic, English, Chinese and Urdu) we asked a total of 15 participants (3 for English and 4 for the other languages) to carry out the same task as the non-experts. All expert and non-expert participants (whether MTurk workers or direct-contact) used the same user interface (Figure 1) as described in section 3.5.

3.4 Experimental Design

Obtaining reliable results from the crowd remains a challenging task (Kazai et al., 2009), which requires a careful experimental design and preselection of crowdsourcers. In our experiments, we worked on minimising the effort required by participants through designing a user-friendly interface.[7] Aside from copying the final output code to a text-box everything else is done using mouse clicks. Poorly designed experiments can negatively affect the quality of the results conducted by MTurk workers (Downs et al., 2010; Welinder et al., 2010; Kazai, 2011).

Feedback from a short sample run with local testers helped us update the interface and provide more information to make the task efficient. Figure 1 shows a sample task for the word ‘car’. The majority of the testers were able to complete the task within five minutes. In response to feedback by some of the testers we provided the “Instructions” section in the six languages under consideration.



Please label the word below with a number of tags that represent its meaning. Attach as many (no more than 10), or as few, as you deem appropriate for **all** senses of the word, placing them in descending order of importance. Tags can be assigned in a (positive or negative manner (for example, "occasionally" is tagged negatively for frequency

To assign a tag, click on the add tag button, and select a box from the category selection. This will add an entry in the list, that .can then be sorted so that the most commonly used tag is at the top

Please remove any unrelated tags and make sure you do not exceed 10 tags in total

To help you with identifying common senses of a word, we have provided a number of links to dictionaries, thesauri, and corpora (where you can see real-world usage). The References button will bring these up over the top of other things, so you can still browse the tags

Example: "Jordan" could be : 'Geographical name' or 'Personal name'

Please submit your selections by clicking in the Submit button at the end of the page and wait until you receive a confirmation message where you need to copy the output-code and provide it to us, as in the following example: f962bed5616612c8c7053f6e97e72b129edb4990-b5f5-47ab-8e7e-bfc61c6515ec

Term to Classify

car

Tags

Please add tags to describe the term, arranging them from the most commonly used sense to the least.

- Movement/transportation: land
- American football

Figure 1: Sample Task for the word “Car”

3.5 Online Semantic Labeling

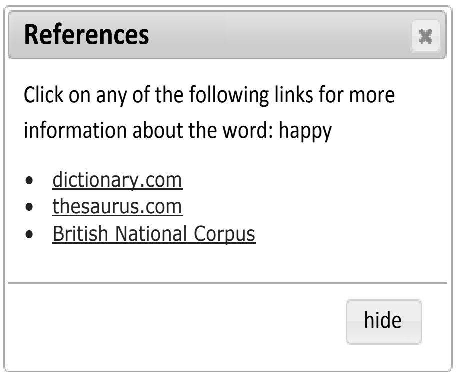
As shown in Figure 1, we asked the participants to label each word presented to them with a number

[4] http://www.mturk.com

[5] https://www.amazon.co.uk

[6] http://www.apple.com

[7] Our underlying code is freely available on GitHub at https://github.com/UCREL/BLC-WSD-Frontend



References

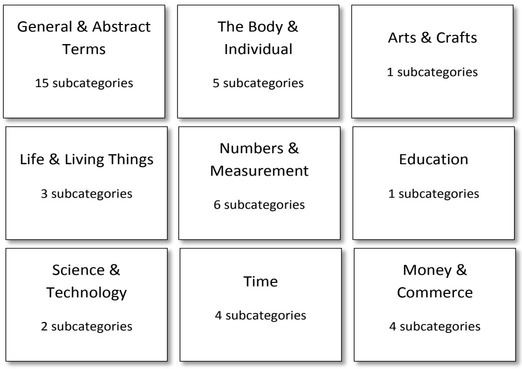
Click on any of the following links for more information about the word: happy

- dictionary.com
- thesaurus.com
- British National Corpus

Figure 2: Dictionary and Thesauri References

of tags that represent the word's possible meanings. The participants were asked to attach as many, or as few, as they deemed appropriate for all senses of the word, placing them in descending order of likelihood.

To assign a tag, the participants click on the `Add Tag` button, and navigate to a box from the category selection where they can select a subcategory (Figures 3 and 4). By following these steps the participants add an entry in the list, that can then be sorted by dragging and dropping the selected tags so that the most commonly used tag is at the top. We asked the users to remove any unrelated tags and make sure they do not exceed 10 tags in total.



General & Abstract Terms
15 subcategories

The Body & Individual
5 subcategories

Arts & Crafts
1 subcategories

Life & Living Things
3 subcategories

Numbers & Measurement
6 subcategories

Education
1 subcategories

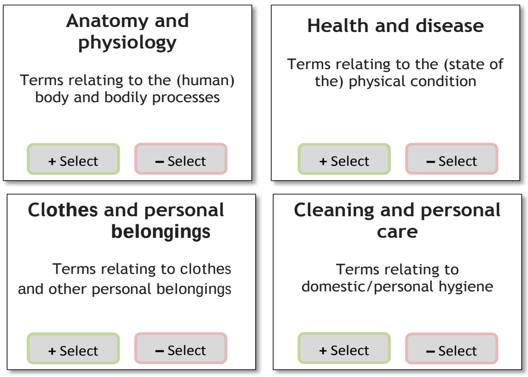
Science & Technology
2 subcategories

Time
4 subcategories

Money & Commerce
4 subcategories

Figure 3: Categories

To help them when identifying common senses of a given word, we provided a number of links to dictionaries, thesauri, and corpora (where they can see real-world usage) for each language. The References are displayed alongside the interface, so they can still browse the tags (Figure 2). Participants are free to use other resources as they see fit. The participants then needed to submit their selections by clicking the `Submit` button at the end of the page and wait until they receive a confirmation message where they need to copy the output-code and provide it to us.



Anatomy and physiology

Terms relating to the (human) body and bodily processes

Health and disease

Terms relating to the (state of the) physical condition

Clothes and personal belongings

Terms relating to clothes and other personal belongings

Cleaning and personal care

Terms relating to domestic/personal hygiene

Figure 4: Subcategories

For each word we targeted a total of four non-expert participants and four expert participants to allow measurement and comparison of the agreement within each group to investigate the variability of task results and participants, rather than to take a simple weighted combination to produce an agreed list.

3.6 Filtering

Even though crowdsourcing has been shown to be effective in achieving expert quality for a variety of NLP tasks (Snow et al., 2008; Callison-Burch, 2009), we still needed to filter out workers who were not taking the task seriously or were attempting to manipulate the system for personal gain (spamming).

In order to avoid these spamming crowdsourced results, we designed a novel task-specific two stage filtering process that we considered more suitable for this type of task than previous filtering approaches. Our two stage process encompasses filters that are appropriate for experts and non-experts, and is applicable whether participants are using MTurk or not.

In stage one filtering, we asked the MTurk workers to select the correct synonym of the presented word from a list of noisy candidates in order to avoid rejection of their HITs. The list contained four words where only one word correctly fitted as a synonym. In order to set up the first filtering task for MTurk workers (on English, Chinese and Portuguese tasks), we used Multilingual WordNet to obtain the most common synonym for each word. The stage one filtering was not needed for Arabic, Italian and Urdu, since these non-expert participants were directly contacted and we knew that they were native speakers and would not submit random results. The synonyms were validated by linguists in each of the three languages and choices were randomly shuffled before being presented to the workers. Stage one filtering removed 12% of English HITS, 2% of Chinese, and 6% of the Portuguese submissions. In our results presented below, we only considered tasks by workers who chose the correct synonyms and rejected the others.

For the stage two filter, we injected random erroneous senses for each of the 250 words into the initial list of tags and the participants were expected to remove these in order to pass. We deliberately injected wrong and unrelated semantic tags in between 'potentially' correct ones before shuffling the order of the tags. For example, examining the pre-selected tags for the word 'car' in Figure 1 we can see that the semantic tag 'American Football' is unrelated to the word 'car' and in fact does not exist in the USAS semantic taxonomy. The potentially correct tag such as 'Movement/transportation: land' does exist in the semantic lexicon. Results where participants fail to pass stage two are still retained in the experiment and we report on the usefulness of this filter in section 4. All participants (MTurk workers and directly-contacted; experts and non-experts) undertook stage 2 filtering. Our experimental design did not reveal to the participants any details of the two stage filtering process.

4 Results and Discussion

To evaluate the results[8] we adopted three main metrics (inspired by those used in the SemEval procedures): Accuracy, Completeness, and Correlation.

Accuracy: measure the accuracy of the participant's selection of tags (W_{Tags}) by counting the matching tags between the worker's selection and the gold standards (G_{Tags}). To compute Accuracy we divide the number of matching tags by the number of tags selected by the participants.

$$Accuracy = \frac{W_{Tags} \cap G_{Tags}}{W_{Tags}}$$

Completeness: measure the completeness of the participant's selection of tags (W_{Tags}) by finding whether the gold standard tags (G_{Tags}) are completely or partially contained within the worker's selection. To compute Completeness we divide the number of matching tags by the number of gold standard tags.

$$Completeness = \frac{W_{Tags} \cap G_{Tags}}{G_{Tags}}$$

Correlation: To test the similarity of tag selection between workers and gold standards we used Spearman's rank correlation coefficient.

In addition to the three metrics mentioned above we used three factors that work as indicators of the quality of the tagging process:

- *Strict*: Whether worker's tags are identical to the gold standard (same tags in the same order);
- *First Tag Correct*: Whether the first tag selected by the worker matches the first tag in the gold standard;
- *Fuzzy*: Whether tags selected by a worker are contained within the gold standard tags (in any order).

For each language we asked for up to four annotators per word (1,000 HITs per language). For Portuguese, where the participants were *all* from MTurk, we only received 694 HITs even though we paid participants working on Portuguese 50% more than we paid for the English tasks.

Table 2[9] shows the aggregate averages of the non-expert HITs. In total, direct-contact participants and MTurk workers performed well and achieved comparable results to the gold standard in places. Around 50% of the English, Chinese, Urdu, and Portuguese HITs had the correct tags selected with around 15% being identical to the gold standards. In nearly all of the cases, Portuguese workers chose the correct tags although they were in a different order than the gold standard. Arabic participants achieved a high completeness score relative to the gold standard tags, but a close analysis of the results show the participants have suggested more tags than the gold standards.

The results for English suggest that the non-expert workers are consistent as can be observed

[8] All expert and non-expert results are available as CSV files on https://github.com/UCREL/Multilingual-USAS/tree/master/papers/eacl2017_sense_workshop

[9] For this and the following tables, we use: Acc: Accuracy, Com: Completeness, Corr: Spearman's, Err: Erroneous tags, Str: Strict, 1st: first tag correct, Fuz: Fuzzy.

Lang	Acc	Com	Cor	Err	Str	1st	Fuz
En	61.4	69.3	0.38	29%	16%	65%	38%
Ar	55.5	87.1	0.35	8%	8%	55%	19%
Zh	45.2	56.1	0.22	2%	15%	46%	27%
It	45.7	47.9	0.06	31%	7%	38%	22%
Pt	58.5	56.3	0.21	18%	19%	50%	94%
Ur	97.6	91.9	0.51	1%	53%	78%	95%

Table 2: Summary of performance [Non experts]

by looking at the Accuracy (Acc) and Completeness (Com) results. Spearman's correlation (Cor) suggests that the workers correlate with the expert gold standard tags, which is consistent with previous findings that MTurk is effective for a variety of NLP tasks through achieving expert quality (Snow et al., 2008; Callison-Burch, 2009). The majority of the workers matched the first tag correctly (1st) by ordering the tags so the most important (core sense) tag appeared at the top of their selection. The erroneous tags (Err) column shows that many workers did not remove some of the deliberately-wrong tags (see Section 3.5). This reflects the lack of training of the workers, but our checking of the results showed that the erroneous tags were not selected as first choice. Strict (Str) and Fuzzy (Fuz) show that many workers were consistent with the gold standard tags in terms of both tag selection and order. It is worth mentioning that languages differ in terms of ambiguity (e.g. Urdu is less ambiguous than Arabic) which can be observed in the differences between language results.

As mentioned earlier, we did not use MTurk for the Arabic lexicon, due to the lack of Arabic natives speakers on MTurk. Instead, we found four student volunteers who offered to help in semantically tagging the words, again without any training on the tagset. The results show consistent accuracy and completeness. It is worth noting that the Arabic participants obtained higher accuracy and completeness scores by having higher agreement with the gold standard tags. The Arabic language participants selected fewer erroneous tags than the English ones. The majority of the participants got the first tag correct. Arabic participants failed to match the order of tags in the gold standards as reflected by lower correlation. This is expected due to the fact that Arabic is highly inflectional and derivational, which increases ambiguity and presents a challenge to the interpretation of the words (El-Haj et al., 2014). Difficulties in knowing the exact sense of an out of context Arabic word could result in disagreement when it comes to ordering the senses (see Section 3.1).

For the Chinese language, the result table shows that there is a slightly lower correlation between the non-expert workers' tags and the gold standards. Observing the erroneous results column we can see that the workers have made very few mistakes and deleted the random unrelated tags. The Strict and Fuzzy scores suggest the results to be of high quality. The participants managed to get the first tag correct in many cases.

For the Italian language, we sourced four non-expert undergrad student participants who are all native Italian speakers but not familiar with the tagset. The participants' results do not correlate well with the gold standards. As the tags description are all in English the annotators found it difficult to correctly select the senses and had to translate some tags into Italian which could have resulted in shifting the meaning of those tags, to communicate with them we had an Italian/English bilingual linguist as a mediator.

As with Arabic and Italian, for the Urdu language we sourced four non-expert participants who are all native Urdu speakers but not familiar with the tagset. Urdu results show that participants correlate well with the gold standards. We also notice a lower percentage of erroneous tags than other languages. The First Tag and Fuzzy scores suggest the results to be of high quality. The participants also managed to get the first tag correct in many cases. The participants all agreed it was easy to define word senses with the words being less ambiguous compared to other languages. This is shown in the high results achieved when compared to non-experts of the other languages.

We received only 694 HITs for Portuguese tasks on MTurk, which suggests there are fewer Portuguese speakers compared to English and Chinese speakers among the MTurk workers. The results for Portuguese in some cases are of very high quality. It should be noted that the gold standard tags were selected and manually checked by a Brazilian Portuguese native speaker expert. There is a difference between European and Brazilian Portuguese which could result in ambiguous words for speakers from the two regions (Frota and Vigário, 2001).

Table 3 shows the results obtained by using the second filtering mechanism to discard HITs where

Lang	Acc	Com	Cor
En	70.4	69.0	0.36
Ar	56.6	87.6	0.34
Zh	45.6	55.9	0.22
It	54.4	53.5	0.09
Pt	61.3	54.0	0.20
Ur	97.6	91.9	0.51

Table 3: Summary of performance with Second Filter [Non experts]

random erroneous tags were not completely removed. This enables us to increase accuracy for English by 9.0%, Italian by 8.7% and Portuguese by 2.8% without negatively affecting completeness or correlation.

In order to allow better interpretation of the non-experts' scores, we repeated the experiments on a smaller scale with up to four experts per language (English, Arabic, Chinese and Urdu), who were already familiar with the USAS taxonomy and were researchers in the fields of corpus or computational linguistics. Experts used the same task interface to assign senses to 50 words each. The results are presented in Table 4. Most notably, experts consistently excel at removing erroneous tags, leaving only a very small number in the data.

English experts performed much better than English non-experts on completeness, correlation and strict measures while their accuracy scores are comparable. Arabic experts performed much better than Arabic non-experts on the accuracy, strict and fuzzy scores while the 1st score is comparable. Chinese experts performed slightly better than Chinese non-experts on Accuracy, completeness and Fuzzy while other scores were comparable. Urdu experts scored relatively more highly on strict and 1st measures while other scores were comparable to Urdu non-experts. Finally, Tables 5 and 6 show the Observed Agreement (OA), Fleiss' Kappa and Krippendorff's alpha scores for the inter-rater agreement between Expert and Non Expert participants. According to (Landis and Koch, 1977) our inter-rater scores show fair agreement between annotators. This serves to illustrate the task is complex even for experts.

Overall these results show that untrained crowdsourcing workers can produce results that are comparable to those of experts when performing semantic annotation tasks. Directly-contacted and MTurk workers achieved similar levels of results overall. This shows that the novel two-phase filtering method used in our experiment is effective for maintaining the quality of the results.

Lang	Acc	Com	Cor	Err	Str	1st	Fuz
En	66.1	83	0.61	1%	31%	75%	40%
Ar	78.8	72.4	0.22	1%	39%	51%	73%
Zh	50.4	60.2	0.21	1%	15%	44%	31%
Ur	96.2	94.8	0.69	1%	63%	89%	93%

Table 4: Summary of performance [Experts]

Language	Measure	OA	Fleiss	K–alpha
English	First Tag	0.82	0.46	0.46
	Fuzzy	0.64	0.27	0.27
	Strict	0.69	0.32	0.32
Arabic	First Tag	0.77	0.55	0.55
	Fuzzy	0.84	0.59	0.59
	Strict	0.21	0.55	0.55
Chinese	First Tag	0.62	0.23	0.24
	Fuzzy	0.75	0.41	0.41
	Strict	0.83	0.31	0.32
Urdu	First Tag	0.83	0.10	0.10
	Fuzzy	0.91	0.35	0.35
	Strict	0.71	0.37	0.38

Table 5: Total Inter-rater agreement [Experts].

5 Conclusion and Future Work

In order to accelerate the task of creating multilingual semantic lexicons with coarse-grained word senses using a common multilingual semantic representation scheme, we employed non-expert native speakers via MTurk who were not trained with the semantic taxonomy. Overall, the non-expert participants semantically tagged 250 words in each of six languages: Arabic, Chinese, English, Italian, Portuguese and Urdu. We analysed the results using a number of metrics to consider the correct likelihood order of tags relative to a gold-standard, along with correct removal of random erroneous semantic tags, and completeness of tag lists. Crowdsourcing has been applied successfully for other NLP tasks in previous research, and we build on previous success in WSD tasks in three ways. Firstly, we have specific requirements for semantic tagging purposes in terms of placing coarse-grained senses into a semantic taxonomy rather than stand-alone definitions. Hence, our experimental set-up allows us to validate the sense inventory in a multilingual setting, carried out here for six languages. Secondly, we extend the usual

Language	Measure	OA	Fleiss	K–alpha
English	First Tag	0.71	0.36	0.36
	Fuzzy	0.58	0.11	0.11
	Strict	0.79	0.20	0.20
Arabic	First Tag	0.66	0.32	0.32
	Fuzzy	0.71	0.05	0.05
	Strict	0.86	0.03	0.03
Chinese	First Tag	0.73	0.45	0.45
	Fuzzy	0.74	0.38	0.38
	Strict	0.85	0.41	0.41
Italian	First Tag	0.67	0.31	0.31
	Fuzzy	0.67	0.03	0.03
	Strict	0.89	0.12	0.13
Portuguese	First Tag	0.64	0.22	0.22
	Fuzzy	0.63	0.13	0.13
	Strict	0.80	0.18	0.18
Urdu	First Tag	0.72	0.16	0.16
	Fuzzy	0.95	0.45	0.45
	Strict	0.74	0.49	0.49

Table 6: Total Inter-rater agreement [Non Experts].

classification task of putting a word into one of an existing list of senses, instead asking participants to list all possible senses that a word could take in different contexts. Thirdly, we have deployed a novel two-stage filtering approach which has been shown to improve the quality of our results by filtering out spam responses using a simple synonym recognition task as well as HITs removing random erroneous tags. Our experiment suggests that the crowdsourcing process can produce results of good quality and is comparable to the work done by expert linguists. We showed that it is possible for native speakers to apply the hierarchical semantic taxonomy without prior training by the application of a graphical browsing interface to assist selection and annotation process.

In the future, we will apply the method on a larger scale to the full semantic lexicons including multiword expressions, which are important for contextual semantic disambiguation. We will also investigate whether adaptations to our method are required to include more languages such as Czech, Malay and Spanish. In order to pursue the work beyond the existing languages in the USAS system, we will extend bootstrapping methods reported in Piao et al. (2015) with vector-based techniques and evaluate their appropriateness for multiple languages. Finally, we will test whether (a) provision of words in context through concordances, (b) prototypical examples for each semantic tag, or (c) semantic tag labels in the same language as the task word, as part of the resources available to participants would further enhance the accuracy of the crowdsourcing annotation process.

Acknowledgements

The crowdsourcing framework and interface was originally produced in an EPSRC Impact Acceleration Award at Lancaster University in 2013. This work has also been supported by the UK ESRC/AHRC-funded CorCenCC (The National Corpus of Contemporary Welsh) Project (ref. ES/M011348/1), the ESRC-funded Centre for Corpus Approaches to Social Science (ref. ES/K002155/1), and by the UCREL research centre, Lancaster University, UK via Wmatrix software licence fees. We would also like to express our thanks to the many expert and non-expert participants who were also involved in the experiments including: Dawn Archer (Manchester Metropolitan University, UK), Francesca Bianchi (Universita del Salento, Italy), Carmen Dayrell (Lancaster University, UK), Xiaopeng Hu (China Center for Information Industry Development, CCID, China), Olga Mudraya (Right Word Up, UK), Sheryl Prentice (Lancaster University, UK), Yuan Qi (China Center for Information Industry Development, CCID, China), Jawad Shafi (COMSATS Institute of Information technology, Pakistan), and May Wong (University of Hong Kong, Hong Kong, China).

References

Ahmet Aker, Mahmoud El-Haj, Udo Kruschwitz, and M-Dyaa Albakour. 2012. Assessing Crowdsourcing Quality through Objective Tasks. In *8th Language Resources and Evaluation Conference*, Istanbul, Turkey. LREC 2012.

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203, Los Angeles, USA. Association for Computational Linguistics.

Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment. In *SIGIR '09: Workshop on The Future of IR Evaluation*, Boston, USA.

Vikas Bhardwaj, Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 47–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 286–295, Singapore. Association for Computational Linguistics.

Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 259–268, New York, NY, USA. ACM.

Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2399–2402, New York, NY, USA. ACM.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using Mechanical Turk to Create a Corpus of Arabic Summaries. In *Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages workshop held in conjunction with the 7th International Language Resources and Evaluation Conference (LREC 2010).*, pages 36–39, Valletta, Malta. LREC 2010.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2014. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, pages 1–32.

Sónia Frota and Marina Vigário. 2001. On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. pages 247–275.

Adam Kapelner, Krishna Kaliannan, H.Andrew Schwartz, Lyle Ungar, and Dean Foster. 2012. New insights from coarse word sense disambiguation in the crowd. In *Proceedings of COLING 2012: Posters*, pages 539–548, Mumbai, India. The COLING 2012 Organizing Committee.

Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. 2009. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 452–459, New York, NY, USA. ACM.

Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In Paul Clough, Colum Foley, Cathal Gurrin, Gareth J.F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 165–176. Springer Berlin Heidelberg.

Adam Kilgarriff. 1997. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA. ACM.

J.Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Laura Löfberg, Scott Piao, Asko Nykanen, Krista Varantola, Paul Rayson, and Jukka-Pekka Juntunen. 2005. A semantic tagger for the Finnish language. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK.

Tom McArthur. 1981. *Longman Lexicon of Contemporary English*. Longman, London, UK.

Olga Mudraya, Bogdan Babych, Scott Piao, Paul Rayson, and Andrew Wilson. 2006. Developing a Russian semantic tagger for automatic semantic annotation. In *Proceedings of Corpus Linguistics 2006*, pages 290–297, St. Petersburg, Russia.

Rebecca Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of North American Chapter of the Association for Computational Linguistics – Human Language Technologies Conference*, Denver, USA. (NAACL HLT 2015).

Yufang Qian and Scott Piao. 2009. The development of a semantic annotation scheme for chinese kinship. *Corpora*, 4(2):189–208.

Paul Rayson, Dawn Archer, Scott Piao, and Anthony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Semantic Labelling for NLP tasks workshop*, pages 7–12, Lisbon, Portugal.

Aanna Rumshisky, Nick Botchan, Sophie Kushkuley, and James Pustejovsky. 2012. Word sense inventories by non-experts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 4055–4059, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Alexander. Sorokin and David Forsyth. 2008. Utility data annotation with amazon mechanical turk. In *In IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

Jean Véronis. 2001. Sense tagging: does it make sense? In *Proceedings of Corpus Linguistics 2001*, Lancaster, UK. UCREL.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2424–2432.

Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. 2009. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA. ACM.