Towards Building Ontologies with the Wisdom of the Crowd

Towards Building Ontologies with the Wisdom of the Crowd

Paula Chocron[1] and **Dagmar Gromann**[2] and **Francisco José Quesada Real**[3]

Abstract. Crowdsourcing provides a valuable source of input that reflects the human diversity of domain knowledge. It has increasingly been used in ontology engineering and evaluation, however, few approaches consider different types of crowdsourcing for data acquisition. In this paper, we compare two crowdsourcing techniques - a mechanized labor-based task and a game-based approach - to acquire shared knowledge from which we semi-automatically build an ontology. This paper focuses on the first two steps of ontology engineering, the forming of concepts and their hierarchical relations. To this end, we adapt a distributional semantic and class-based word sense disambiguation approach and a knowledge-intensive tree traversal algorithm. Each step along the process and the final resources are evaluated manually and by a gold standard created from Wikipedia data. Our results show that the ontology resulting from data obtained with the mechanized labor-based approach provides a higher level of granularity than the game-based one. However, the latter is faster and seems more enticing to participants.

1 INTRODUCTION

Creating knowledge resources manually is a time- and cost-intensive task [26], and the resulting resources are in general difficult to maintain. Moreover, when resources are created by individuals (*experts* in the domain and the technique), in many cases they are not free from arbitrariness. The default alternative to manually crafting knowledge resources is to develop techniques that automate the process, or at least parts of it. The ontology learning community has developed different automated approaches, using tools that range from machine learning [16] to NLP-intensive approaches [22]. These methods extract information from either a structured (e.g. WordNet) or unstructured (e.g. text) existing corpus, and are therefore strongly dependent on the existence and quality of such a corpus. As an alternative, and paired with a general growing interest in these kind of techniques, in the past years the community has proposed different applications of crowdsourcing methods to ontology engineering (e.g. [7, 15, 28]). We contribute to this community effort by comparing two distinct crowdsourcing approaches to the task of knowledge acquisition for building ontologies semi-automatically.

Crowdsourcing is a problem-solving method that relies on a collective of non-experts (a *crowd*) performing short and accessible tasks that are then combined to tackle a larger problem. Crowdsourcing methods are particularly well suited for tasks that are difficult to automatize completely, but are at the same time too large to be completed by just one person, or that benefit from the diversity of the participants, as is the case with our approach. This includes, for example, many information retrieval or classification tasks, often in complex human domains, such as natural language. The question of how to increase the attractiveness of crowdsourcing methods to make the participation more appealing has received much attention as of late. While one way is to provide explicit, in general monetary, incentives, other methods rely on intrinsic rewards, such as learning a language [23], helping a cause, or having fun. This last category is particularly exploited via the *Games With a Purpose* approach [21, 24].

This paper proposes an ontology learning technique that combines crowdsourcing to retrieve data with automated methods to organize it. Instead of crowdsourcing the ontology building process as it is frequently done, we leverage diversity by crowdsourcing the data acquisition step. Thereby, we obtain domain knoweldge that reflects the human diversity of domain knowledge and brings ontologies closer to their initial aim of representing shared knowledge. We build two ontologies from scratch using the data obtained from two separte crowdsourcing methods and then compare them to each other as well as to a third gold standard ontology obtained from Wikipedia data. While this knowledge production technique has all the advantages of collaborative methods, the obtained data is usually not organised, which represents a technical challenge when building an ontology with it. Thus, we implement and compare different methods to disambiguate the retrieved data categories and we build a taxonomical structure with it.

We focus on the task of building an ontology for a particular concept, identifying all the related categories that could be used when describing an instance. We chose to perform our experiments using the concept of *city*, mainly for three reasons. First, it is a topic with which the crowds are in general familiar. Second, it belongs to a category of particularly fuzzy, collectively constructed concepts, which makes it ideal to be crowdsourced. Third, a sound representation of city has become something particularly necessary in the last years, with the growing interest in visions such as the one of Smart Cities [3]. The ability of a city to share and re-use data has become a key indicator for a Smart City and a domain ontology that contains categories typically characterizing a city can facilitate this task as well as the integration of data across Smart Cities.

To obtain these ontologies, we first implement two crowdsourcing methods (a direct and a game-based one) in which we ask participants to describe instances of a city on $CrowdFlower$[4] and in a game we developed. We consider this kind of crowdsourcing *implicit*, since participants have to solve a different problem from which the desired data are then extracted in a post-processing phase. An *explicit*

[1] Artificial Intelligence Research Institute (IIIA-CSIC) and Universitat Autònoma de Barcelona, email: pchocron@ iiia.csic.es

[2] Artificial Intelligence Research Institute (IIIA-CSIC), email: dgromann@iiia.csic.es

[3] University of Edinburgh, email: fquesada@inf.ed.ac.uk

[4] `https://crowdflower.com/`

approach would consist in asking people for characterizations of the general concept of city itself. Implicit crowdsourcing techniques are useful in order to make the task more attractive, fun, or *gamifiable* than the explicit approach. We also believe that it can lead to richer and more fine-grained ontologies than the explicit one. However, the direct comparison between explicit and implicit crowdsourcing is yet to follow. The kind of techniques we propose here is particularly applicable when describing abstract concepts that do not have a clear physical correspondence, where the properties are less evident.

Our post-(crowdsourcing)-processing phase consists in extracting categories related to cities from the crowdsourced description by analyzing the obtained natural language expressions. To this end, we first disambiguate the senses of these expressions, for which we implement two techniques - a distributional semantics and a class-based approach. We also consider the next step in ontology building, which is adding a taxonomical backbone to the resource by relating the disambiguated concepts hierarchically and extracting their superordinate classes. Finally, we evaluated our approach by comparing its results to an existing, also crowdsourced, description of cities that we extract from the Wikipedia Tables of Contents (TOCs) of individual city pages.

After discussing related work, we describe our approach following the traditional structure of method, results, and discussion. We first explain the techniques that were implemented for each step, then present the results obtained with each of them, and finally compare them and discuss advantages and drawbacks of each one. In the last section we present future work and some concluding remarks.

2 RELATED WORK

Due to the difficulties that the manual crafting of ontologies present, the field of ontology learning has been extensively studied in the past years [12]. Many of these approaches, particularly those in the first years of the area's development, rely on predefined patterns and rules or static resources, such as WordNet [26]. However, these static approaches have two drawbacks, namely they are neither scalable nor easily portable between domains. Recent approaches seek to be more dynamic, for example by using machine learning to extract relations from an existing seed ontology [16] or to develop axioms extracted from text [22].

Using static resources in ontology learning is not straightforward due to the multiplicity of senses associated with each word. To address this problem, Bentivogli et al. [2] associate senses with a WordNet domain ontology they create and which we also use herein to classify words. A similar idea is presented by [8] who associate the Kyoto ontology of the project with WordNet senses and also a number of upper level ontologies. Those associations are then used to present a class-based word sense disambiguation method we adapt in this paper. Alternatively, distributional semantic approaches have been investigated for word sense disambiguation with context-poor data sets. For instance, Basile et al. [1] extract DBpedia glosses for each word in tweets and then compute the cosine similarity between the context of the word in the tweet and each gloss to find the most related one(s), a second approach we adapt in this paper. Similarity between sets of words can be computed by composing their vectors in different ways; in [1] the authors use addition.

The use of crowdsourcing techniques has received considerable attention across research fields in the past few years [27]. For instance, crowdsourcing is highly popular in Natural Language Processing (NLP), such as for named entity recogntition [9]. In ontology learning and building, crowdsourcing has mainly been used in an explicit fashion, asking users to relate concepts hierarchically [6] or evaluate already learned relations and term clusters [7]. Additionally, it has been used as a method to align ontologies with each other [17, 20]. Most frequently, crowdsourcing has been applied to ontology evaluation both for verifying subsumption hierarchies [15] as well as entire ontology statements [28].

Among these crowdsourcing techniques, games are particularly important since they offer an interesting way to motivate humans to solve large-scale problems that are currently beyond the ability of computers [18, 24]. Some well-known examples are Duolingo [23], an approach to crowdsourcing the translation of the Web, and reCAPTCHA [25], a method for digitizing paper copies of documents. Approaches that use 'Games with a Purpose' build on the instrinsic motivation of participants to learn something new. For instance, Dumitrache et al. [5] use gamification and crowdsourcing to create a gold standard for annotations of medical texts. Luengo-Oroz et al. [11] develop a game for counting malaria parasites in images of thick blood films, while Deng et al. [4] and Zou et al. [29] focus on feature discovery and image categorization. Individual ontology engineering tasks have been crowdsourced as games as well, such as for classification and population [19]. In [14] a game is proposed to obtain attributes for concept descriptions. Their approach is explicit in that it asks players to name properties directly. In combination with ontologies, a specific part of the ontology building task is usually crowdsourced but not the knowledge acquisition step that precedes the ontology building as in our approach.

3 METHOD

In this paper we present a method to build ontologies for the concept of *city* from data obtained with crowdsourcing techniques. We use two different implicit crowdsourcing methods, in which we ask participants to describe specific instances of cities as direct question and in a game to obtain a general characterization of *city* as a general concept. We consider *city* to be a particularly good concept to perform this experiment, since it has clear instances which are in general well-known by a random crowd. In addition, although a city can be uniquely identified by means of its coordinates, these are in general not the most immediate characteristics that come to mind, and the resources used when describing an instance are very varied.

From the descriptions obtained with the crowdsourcing methods, we extract general categories on which we build a hierarchical taxonomy to obtain a preliminary ontology for the concept of *city*. We consider the results obtained to be seed ontologies that can be used for further ontology learning rather than fully formalized ontologies; nevertheless, they can be seen as a schema of a city characterization. To evaluate this claim, we compare them to a gold standard ontology that we manually and collaboratively build from Wikipedia TOCs of pages describing specific instances of cities, countries, regions, and continents. The complete process of our approach is depicted in Figure 1, where rectangles are steps and circles are the different techniques that we explore.

3.1 Data Collection

Our method for collecting data by means of crowdsourcing can be subdivided into two separate techniques: (1) mechanized labor-based knowledge acquisition, and (2) game-based knowledge acquisition. Mechanized labor refers to popular crowdsourcing platforms where people complete mechanical tasks in exchange for monetary rewards, e.g. *CrowdFlower* or *Mechanical Turk*. In this type of data collection

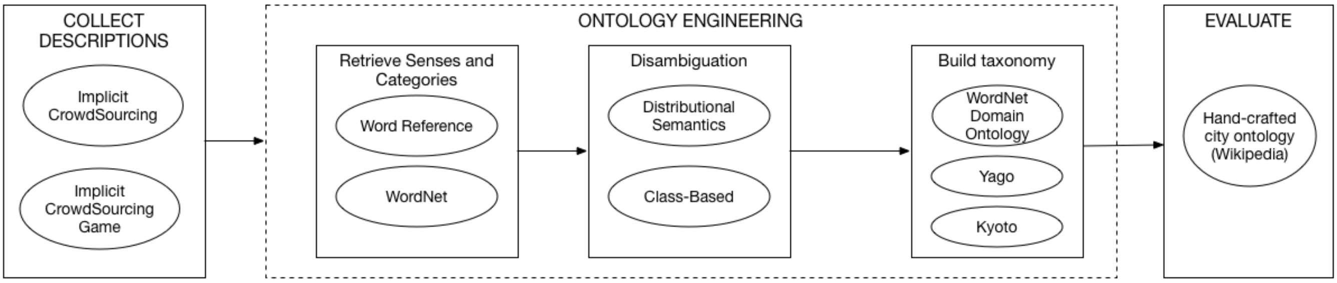

Figure 1: Steps followed in the ontology building procedure

method participants were asked to provide the first ten words they associate with a city name displayed to them. In contrast, in a game-based elicitation of knowledge participants provide the desired information while playing a game without being asked direct questions. We utilized a list of 300 city instances derived from online listings of popular cities that were retrieved by a search engine query. In both tasks it was possible to skip to the next city if a participant was not familiar with a specific city instance.

Both tasks focus on the collection of common nouns in combination with verbs and adjectives. There are two main reasons for this restriction: (1) proper nouns trivialize the identification of cities as they uniquely identify them, e.g. Eiffel for Paris, and (2) we were interested in ontology building from common language and not based on instances or named entities. In both types of activities participants were explicitly instructed to comply with this input restriction. Additionally, measures were taken in both tasks to enforce this restriction and the non-conforming characterizations were omitted from the final data set. The results from the first data set are distinct from the second data set since the nature of the game required us to provide descriptions of the city obtained from the first technique as input for the game.

3.1.1 Mechanized Labor-Based Knowledge Acquisition

To obtain first characterizations of cities, we uploaded the list of 300 cities to the crowdsourcing platform *CrowdFlower*. Questions presented to the crowd provided the name of the city, its country name, latitude and longitude, ten input fields for city descriptions, and the option to skip to the following city if the city was not known to the worker. In addition, each worker was asked 20 test questions to ensure their ability to comply with the instructions regarding the input restrictions, such as use of a common noun or noun phrases with adjective and verbs, use of loan words but no words that are not English, and omission of personal opinions. For instance, we asked workers whether Breaking Bad is an adequate description of Albuquerque, USA. Since this is the title of a TV series and thus, a named entity, this question had to be negated. The ability to comply with instructions was also tested by using misleading descriptions, such as the description of Liverpool with U2. Each test question was equipped with a detailed explanation for the correct answer so that participants who did not fully read or understand the instructions were prepared for the actual question of the task.

For quality assurance four measures were taken: (1) the actual run was preceded by a test run, (2) each worker was asked twenty test questions, (3) only workers with an accuracy exceeding 70% on the test questions could participate in the task, and (4) only workers who spend more than ten seconds on each question apart from the test questions would be considered. Furthermore, we limited this task to workers with English as their first language since we required an English data set and such word association tasks are difficult in a second language. An initial test run with a subset of the cities helped evaluate the kind of results we were to expect and modify the test questions and project settings based on the feedback from the crowd. In fact, those modifications strongly improved the quality of the results as well as the time needed to obtain them in the actual second run.

Obtained city characterizations were deduplicated automatically by applying similarity measures from the WordNet Similarity for Java (WS4J) library[5] combined with the Levenshtein distance [10]. On this basis the most frequently provided and deduplicated city descriptions were identified and then evaluated manually.

3.1.2 Game-Based Knowledge Acquisition

In this second crowdsourcing technique, participants played a Taboo game of cities adapted from the popular board game *Taboo*. There are two roles a player might assume: describer and guesser. The describer provides hints to the guesser that describe a given city and the guesser responds with a city name that is believed to be the correct result. The objective of the game is to obtain the name of the described city from the guesser. As a further restriction, the describer may not use any of the phrases that are provided as taboo words along with the city.

The taboo words of this game were obtained from the first data collection method. Thereby, it was ensured that there is no overlap between the data set gathered with the first collection method and this second crowdsourcing method. Additionally, in order to play this game, Taboo words are needed. For the hints, the same conditions as in the first method were applied for the same reasons. This meant that we needed to limit the type of hints people provide when playing the game.

Players were recruited at the University of Edinburgh by means of internal mailing lists and personal contacts of our local colleagues within the ESSENCE project [6]. As with the first technique, we restricted the participation to native English speakers. Each participant obtained a small shopping voucher in return for their participation. The number of games per participant was not limited.

To ensure that the input complied with our restrictions and to enable several simultaneous games, we developed an online platform [7] and pre-scheduled game sessions with up to nine players at a time.

[5] `https://code.google.com/archive/p/ws4j/`
[6] `http:\essence-network.com`
[7] `http:\taboo.iiia.csic.es`

The first player to log onto a game would be assigned the guesser role. The second player to join a game session would be the describer, who in contrast to the guesser would see the city name, country, and Taboo words. The game commences by the describer providing a hint and ends with the correct guess from the guesser. Players were newly assigned automatically and anonymously to each game. This should prevent participants from providing clues based on previous experiences in case of acquainted players.

The final data set is limited to successful games that follow the restrictions of the initial instructions. A successful game is one were the city was guessed correctly based on the provided hints. This ensures the quality of the hints, i.e., they are indeed associated with the city being described to a degree that allows a human player to identify the city. Naturally, there might be many reasons for the inability of a guesser to provide the correct city name, which, however, we did not investigate for this paper and instead relied on the quality-assured hints of successful games.

3.2 Ontology Engineering

The task of building ontologies, known as ontology engineering, is commonly divided into four major steps that can be implemented with different engineering methods:

1. concept formation
2. concept hierarchy building
3. building non-taxonomic relations
4. axiom discovery

This list is non-exhaustive, and some approaches also include, for instance, ontology population as another step. In this section, we present the methods we implemented for building ontologies for the *city* concept from the data sets that resulted from the two crowdsourcing methods described previously. In this paper we focus on the first two steps of ontology engineering: forming the concepts and building a hierarchy. As we discuss later, the third step can be initiated with the methods we use but will be the subject of another paper since we do not evaluate it here.

Concept formation refers to the process of clustering terms based on their more general categories. Thus, for this step we required the general concepts related to *city* that were represented by the descriptions of city instances. For example, if sun was related to Barcelona, we wanted to extract weather as a general characteristic of a city. To this end, we retrieved the available senses and classifications for each noun and noun phrase from WordNet and an online dictionary. We noticed that, although these methods return adequate categories, an unexpected level of complexity arises given the multiplicity of senses that exist for each hint. Thus, our method required a step of word sense disambiguation for which we implemented two ideas: (a) a distributional semantic approach taking the city as context, and (b) a class-based approach that does not consider the city. Finally, we present our method to build the taxonomy, extracting more general concepts for the categories obtained. In this section we explain the techniques we used for each of these ontology engineering steps depicted in Figure 1.

3.2.1 Sense Extraction and Classification

The first of our approaches to extract general categories from descriptions of specific cities uses Word Reference [8], an online dictionary that associates words with general labels that can be generally seen as its superordinate class. For example, Sushi is labeled as Food. To use this information, we first extract all nouns in city descriptions and retrieve all existing glosses and categories from Word Reference. A second approach consists in using WordNet to obtain the categories. Due to the fine-granular nature of WordNet senses, it was necessary to use ontologies associated with WordNet synsets to obtain general categories, as described in detail below.

3.2.2 Sense Disambiguation

With context-poor and highly ambiguous input data, word sense disambiguation for the purpose of term clustering and concept formation is a highly challenging task. At times the disambiguation is not even easy for human users, e.g. curse for Cairo could relate to a film, urban legends, or verbal expressions. Both data sets derived from the described crowdsourcing techniques consist of single common nouns or noun phrases with their associated city name as the only context. To address this challenge, we tested two different approaches to word sense disambiguation: (1) a distributional semantic approach, and (2) a class-based approach. In both cases the objective is the identification of the sense that is most closely related to a city, which is then used to form ontology concepts.

All initial input data were submitted to an NLP preprocessing step to identify all common nouns in the data set and lemmatize them. For this we used the NLTK[9] in Python for the distributional approach and the Stanford CoreNLP library[10] in Java for the class-based approach for no reason other than the personal preference of the developers. The former used individual tokens only, while the latter approach first queried noun phrases. If the noun phrases returned no result, the head noun of the phrase was identified by CoreNLP and submitted to the sense query component.

Distributional Semantics-Based Disambiguation Due to the nature of our data there is no real context for the words used. Therefore, our approach consists in computing the similarity of each definition of a word extracted from the lexical resources with the vector of the city. For example, if Paris was described with love, for which we retrieved three definitions, we compute the vector for each definition and their similarity with the vector for Paris, and chose the one with the highest score. After some initial experiments we combined the vector of the city with the vector for each data element from the crowdsourcing techniques since it substantially improved the disambiguation of the word's senses. For instance, in the example above we would combine Paris and love and then compare the result to the three glosses retrieved from the lexical resource. Instead of DBPedia, we opted for an extraction of senses from Word Reference and WordNet since it is faster and less noisy. Furthermore, the categories retrieved along with the senses in Word Reference seemed promising for the classification task. We implemented two ways of composing vectors: the addition used in [1] and a simple average of individual vectors, that is the standard way to compute similarity between sets of words in the word2vec Python package. We chose this last option after performing a general initial comparison.

Class-Based Word Sense Disambiguation To follow up on a second idea, we investigated a class-based sense disambiguation approach adapted from [8]. Although the use of WordNet to disambiguate words is wide-spread, one of the major issues is the high

[8] http://www.wordreference.com

[9] `http://www.nltk.org/`

[10] `http://stanfordnlp.github.io/CoreNLP/`

granularity of its senses. For instance, querying architecture returns five distinct senses ranging from architecture as a profession to computer architecture. One method to alleviate this situation is the semantic classification of WordNet senses by using associated ontologies. The approach in [8] associates WordNet senses semi-automatically with the ontology Kyoto[11].

In this three-step algorithm, we first extract all senses associated with an input noun or phrase from YAGO[12] and query Kyoto for each association with each retrieved sense. In a second step, the algorithm traverses the sense hierarchy in YAGO and searches for categories by again querying Kyoto and searching for WordNet domains associated with individual senses. Since the mapping to WordNet domains is not consistent in YAGO, each sense label queries the WordNet domain ontology for string matches and adds them to the resulting collection of categories and senses. The third step consists of extracting all tokens from each label of a category and ranking them according to frequency. To find the best sense, the most frequent word of all senses and the previously evaluated Word Reference categories from the distributional semantics approach are utilized as determining factor on which sense to return. The extracted and evaluated Word Reference category is added as an additional weight to the decision of which category to chose as the final one and the same approach could be done without this additional weight. Queries to WordNet that immediately return a WordNet domain along with the senses are not submitted to this process but instead classified by the domain directly.

3.2.3 Taxonomy Building

In order to build a hierarchical backbone for an ontology, we query YAGO, WordNet domains, and Kyoto relations. Although some of the upper ontologies in Kyoto are highly useful, we exclude DOLCE since it is too abstract for our purpose, namely building a resource that represents categories associated with the general concept of *city*. The senses we obtain from the word sense disambiguation tasks are utilized to retrieve the Kyoto concepts and WordNet domains directly associated with the sense. Additionally, we traverse the YAGO sense hierarchy up two levels to obtain all senses and domains associated with the disambiguated and evaluated sense. If the word or sense is directly associated with a WordNet domain we extract the superordinate level of the respective domain in the WordNet domain ontology where available. The WordNet domain ontology currently only provides one hierarchical level associating domains with their more general level. If there is no WordNet domain we query Kyoto and extract all concepts that are associated with a sense by means of a subclass relation. The focus here due to the data set is on nouns, which is why we do not extract any concepts related to verbs or adjectives. In case this step returns several concepts, we manually select the best hierarchy for a given WordNet sense.

3.3 Evaluation

Each word sense disambiguation approach is evaluated manually by at least two fluent/first language English speakers. For WordNet, the senses were rated regarding their correct specification of the input description as either *correct* or *incorrect*. For Word Reference, both the categories and the definitions were rated since the former was used as a weight in the taxonomy building task. Only senses and data on which both raters agreed were submitted to the ontology engineering task. The resulting seed ontologies with a hierarchical backbone were again manually evaluated by two ontology engineers.

In a second evaluation step for the ontologies, we compared them with another crowdsourced classification of concept properties that is obtained from describing city instances: one obtained from the TOCs of Wikipedia. The usefulness of TOCs of Wikipedia for building knowledge resources has been acknowledged before [13]. Each Wikipedia page of a specific city is organised in a tree of sections (for example, dog has the subtree Biology → Anatomy → Size and Weight). These TOCs work naturally as an organisation of categories that are important to describing something. Moreover, although Wikipedia establishes certain patterns that authors should follow, TOCs are mostly originated from a collaborative attempt at describing things in the world, here cities.

To build a general ontology for city descriptions we chose 20 random cities from the list of cities we used for the crowdsourcing and merged the TOCs in their Wikipedia pages, keeping the most general ones. In this way, we removed categories that were very specific to one city or region (such as “2.1.1 Legend of the founding of Rome” for the city of Rome). This was done by four ontology engineers in a collaborative shared task to avoid personal biases. Each created an ontology from five different cities, and then all together collaboratively discussed how to merge them to get a common taxonomy. In general it was easy to achieve an agreement, which suggests a high degree of consistency in Wikipedia’s TOCs.

We repeated the same process for countries, regions, and continents and merged the final result to a four-layered knowledge resource reflecting the four main levels we found in the city descriptions. At times people utilize those levels of granularity to describe a city. For instance, nasal vowel relates to Portuguese and Portugal rather than Lisbon while wall clearly relates to the city of Berlin. However, both are used to describe the respective cities in the crowdsourcing tasks.

4 RESULTS

Results are structured in line with the method section to facilitate their traceability. We first report on the obtained data sets from the two distinct crowdsourcing techniques before we detail the results of the ontology engineering method. The evaluation of the word sense disambiguation methods was done by human users and the resulting seed ontologies from both crowdsourcing data sets were evaluated by using a manually curated gold standard ontology based on Wikipedia TOCs.

4.1 Data Collection

From the CrowdFlower platform we derived a total of 6,238 descriptions for 275 of the 300 cities, 25 not being described by a single user. For 244 cities the number of descriptions exceeded 5, which meant they could be kept for the game of Taboo. Similarity measures and simple string-matching techniques were employed to deduplicate the results and identify the most frequent words from this set. This resulted in 576 descriptions for 226 cities where frequent meant that more than one user provided the same characteristic. For the Taboo game we manually chose several additional salient descriptions as Taboo words, while for this task of ontology building we decided to keep only the most frequent ones as a quality assurance measure. We kept duplicates across the data set but de-duplicated the descriptions

[11] http://weblab.iit.cnr.it/kyoto/xmlgroup.iit.cnr.it/kyoto/index.html

[12] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

of each city since many cities are, for instance, a capital. This was particularly crucial for the distributional word sense disambiguation approach that considers the city as the context for individual words. To ensure the comparability of the two data sets, both contain the same instances of cities and their descriptions and all other city instances that were not described with the second method were also omitted in the first data set for this paper. This drastically reduced our data set to 322 descriptions of 62 cities, with an average of 5.65 descriptions per city.

Each city was described by a total of 12 participants, who had the option to skip a city shown to them in case they were not familiar with it. We obtained a total of 3,616 trusted judgments over six days the task was online, where trusted refers to workers with more than 70% accuracy on the test questions and an answer time exceeding ten seconds for each question. In fact, the trust level for this task was extremely high with 91% on average. The 80 participating workers were mainly based in the United States with 65% followed by Great Britain with 30% and the remaining workers came from New Zealand and Australia. We did not limit the number of descriptions that could be provided by an individual participant and some top contributors provided up to 85 descriptions.

The most frequent input data from this first task and a predetermined number of handpicked most salient properties for each city were utilized as taboo words for the remaining 226 cities. The predetermined number of taboo words per city depended on the number of descriptions provided for each city by the crowd: >25 descriptions = 12 taboo words, 20-25 descriptions = 10 taboo words, <20 descriptions = 8 taboo words. Those benchmarks were based on the assumption that more descriptions in the first task require more taboos in the game since people seem to have more associations with those cities and we wanted to keep the game challenging.

The remaining 226 cities were provided to a total of 30 users in five online sessions on the platform. This resulted in 316 games, of which 174 were successful, i.e., the city was guessed correctly in 174 cases. Those successful games were manually evaluated by 12 ontology engineers and researchers regarding their conformance to the restriction to common nouns and the rules of the game, e.g. not containing a taboo word. Two of those engineers evaluated all selected successful and compliant games in a final quality assurance task, to ensure that all games corresponded to the established quality criteria. This process reduced our data set to 73 games of 62 cities and a total number of 202 descriptions of cities. This set of 202 descriptions of the second technique and the 322 descriptions from the first crowdsourcing technique provided the input to our ontology building method.

4.2 Ontology Engineering

4.2.1 Category Extraction

In general, querying Word Reference (WR) and WordNet (WN) provided definitions for the words in the descriptions, although there were some exceptions, most of which were actually words in foreign languages. From the found words, in most cases the correct sense (the one intended by the describer) was available as a definition, as shown in Table 1. ‘Hints’ refer to data obtained by the game by the describer and ‘taboos’ are the result of the first mechanized labor-based task. By ‘not available’ we mean that the word is in the resource but the required sense is not. In some other cases, the describer used the word in a very complex or informal way, which was not included in our resources. This is the case of, for example, using sack to describe Sacramento.

	Available	Not Available
WN (hints)	112	6
WN (taboos)	199	1
WR (hints)	109	9
WR (taboos)	194	6

Table 1: Availability of correct senses for WordNet and Word Reference

We also measured the number of available correct categories in Word Reference depicted in Table 2. In this case the value is lower, because many glosses in Word Reference are not classified into a category. At times, the categorization in the resource is not entirely accurate, as for instance beer is classified as wine instead of alcoholic beverage. Nevertheless, in most cases the quality of the categories is surprisingly high.

	Available	Not Available
hints	77	42
taboos	132	64

Table 2: Availability of categories in Word Reference

4.2.2 Sense Disambiguation

The distributional semantics approach provided the sense in WordNet that was closely related to the pair $(word, city)$ for each hint. We used only the words of which the intended sense was in WordNet, and classified the results as *correct* or *incorrect*. For the Word Reference data we additionally retrieved the closest sense between the ones that were related with a category, since this would be the category assigned to the word. We followed the same criterion for evaluating the categories. Each of the results was evaluated by two ontology engineers, and we considered as correct the intersection of those in both evaluations. The results of this process are depicted in Table 3, where WN refers to WordNet and WR to Word Reference and only the senses available in the resource have been considered. In brackets we indicate the data set, which is ‘taboos’ for the mechanized labor-based approach and ‘hints’ for descriptions obtained from the game-based crowdsourcing technique.

As can be seen from the summary of the results in Table 3, the F-Measure or accuracy of retrieving the correct WordNet sense for both data collections almost reaches 80%. Given the highly ambiguous nature of the input data and the lack of context, we consider this a good result. For instance, boot has 7 different senses and no obvious connection to Wellington. Our algorithm identifies the correct sense of ‘footwear’, since the description most likely hinted at the famous ‘Wellington rubber boot’.

In this approach every exact duplicate for different cities was kept. We considered the description of one city with the term architecture different from the same string for another city. And in fact our results proved this point since Agra was associated with the profession of designing works of architecture, while Berlin returned the discipline of architecture as a field. This shows that the city vector has an effect on the selection of a sense. While we noticed this fact as part of our study, the number of duplicates in our data set was not sufficient to provide a proper analysis of the impact of the city vector on the sense selection, which is definitely interesting for further experiments.

In contrast to the senses, the retrieved results of the categories depicted in Table 4 from Word Reference were considerably lower,

	Correct	Incorrect	F-Measure
WN (hints)	89	23	0.79
WN (taboos)	164	35	0.82
WR (hints)	74	35	0.68
WR (taboos)	122	72	0.63

Table 3: Sense Evaluation for DS approach in WordNet and Word Reference

since not all glosses were classified in the resource itself. Nevertheless, if categories were available, the quality and accuracy was very high. For instance, star for the city Cannes provided the category 'show business'. In order to retrieve meaningful categories, several restrictions had to be added to the algorithm. Firstly, we decided to ignore all categories that classified language usage, such as 'slang term'. Secondly, since one of our crowdsourcing restriction was use of only common nouns, we omitted all categories that referred to proper nouns of any kind. With those restrictions in place, the retrieval of categories led to an accuracy of more than 84% for both data sets.

	Correct	Incorrect	F-Measure
hints	65	12	0.84
taboos	119	13	0.90

Table 4: Category Evaluation for DS approach in Word Reference

As a class-based approach to disambiguating words, the city of the description is not taken into accounting as no difference in the sense selection could be expected. The approach queries the data in YAGO and Kyoto and returns a result irrespective of the city. Thus, the results presented in Table 5 sum to a different total than the results of the distributional approach provided in Table 3. Furthermore, instead of categories this approach considers WordNet domains (WND) for both types of data sets that are directly associated with data as they are queried in YAGO. This is based on the assumption that such domains provide an excellent basis for disambiguating our city descriptions. The results support this point since all the domains that were directly obtained on the first query were accurate categories for the input data.

One further difference between the distributional semantic and the class-based approach is the type of input data. While the former queries individual words in combination with cities, the latter first attempts to retrieve senses for noun phrases, such as red carpet, and, only if it does not retrieve any result, queries the head noun of each phrase. This head noun identification succeeded in 44 out of 48 cases of compounds in all data sets using the Stanford CoreNLP parser. Failures can be attributed to imperfect input, such as *embargo lift instead of lifted embargo to refer to Havanna or to be precise to Cuba and the U.S. trade embargo that has been recently lifted. The head noun that was queried for this example was *lift* which returned a sense related to skiing. Specific symbols equally constituted a problem for the parser since ex-empire remained unchanged and thus did not return a sense, which would have been achieved by only querying empire.

All results were rated by two experts in a separate task and only the ones that were agreed upon are presented in Table 5. The accuracy for each input depended on the sense that was provided or in case of domains on the domain label as well as its superordinate class. For a total of 27 input phrases the raters did not agree and thus those data are neither considered here nor in the taxonomy building task.

	Correct	Incorrect	F-Measure
WN (hints)	78	19	0.80
WN (taboos)	78	8	0.91
WND (hints)	5	0	1
WND (taboos)	16	0	1

Table 5: Class-Based sense disambguation results

4.2.3 *Taxonomy Building*

When building the hierarchical backbone of the two ontologies for the two different data sets, we utilized the disambiguated senses from the previous task. Our approach consisted of following the sense up the hierarchy for two levels and extracting all `subClassOf` relations from Kyoto. Only for the 21 WordNet domains directly associated with the first query no further disambiguation was necessary since each domain returned exactly one additional hierarchy level subordinate to the domain. For instance, the sense skyscraper_104233124 returned the domain wordnetDomain_building_industry, which is in turn narrower in meaning than the wordnetDomain_architecture. One issue we faced in this regard is the poor coverage of WordNet domains in YAGO, which is why we always performed a string matching of sense labels and WordNet domains, which returned twice as many domains as querying YAGO alone. In the final ontology, the retrieved WordNet domain hierarchy was still evaluated manually against the other hierarchies for accuracy and adequacy.

From Kyoto, we retrieved up to 32 senses on the second level of hierarchy. For instance, `victim` provided mostly `person` on the first level but then explodes on the uppermost level to 32 different types of agents and social figures. This number already excludes DOLCE concepts and senses related to any other part-of-speech type than nouns in Kyoto. Thus, the manual effort involved in deduplicating the retrieved hierarchies is rather high and it is definitely worth investigating automated methods in the future. The first and the second level of hierarchy in Kyoto are frequently identical but still relate by means of a subclass relation. Those were eliminated as well.

One step to further reduce unnecessary complications was to deduplicate hierarchies in the ontologies obtained from identical senses, since we also in this step abstract away from the city and thus this context. This means each sense is only included once across the data set and with one specific hierarchy. This step reduced the number of obtained superclass concepts from 426 to 60 for the hints and from 301 to 48 for the taboo words and phrases.

The differences in level of granularity were kept in this experiment. The following two example hierarchies for animals show this difference: `crocodile` $\sqsubseteq$ `animal_fauna` $\sqsubseteq$ `organism_being` as opposed to `dingo` $\sqsubseteq$ `mammal` $\sqsubseteq$ `animal_fauna`. While the first animal is directly mapped to `animal` and then a general concept of `organism`, the Australian representative is first mapped to `mammal` and would require one more hierarchical level to reach the same level of abstraction as the first.

4.3 Evaluation

This section describes the evaluation of our seed ontologies against each other and the Wikipedia ontologies we created for the general concept of city and for the evaluation of their semantic equivalence, their 'citiness'. First, we evaluate the correctness of the extracted Word Reference categories. Then we compare the two taxonomies with a four-layer ontology extracted from Wikipedia TOCs of cities.

This step serves to evaluate whether the crowdsourcing techniques provided semantic classes that are closely related to descriptions of the general concept city by comparing them to a manually created gold standard, our four-layer Wikipedia ontology describing a city on the city, country, continent, and region level.

4.3.1 Word Reference Categories

We evaluated the Word Reference categories by comparing them with the city ontology that we built from Wikipedia TOCs of specific cities. We performed this evaluation only over the categories that were disambiguated correctly with the distributional semantics approach, since we are interested in how far the data collected by crowdsourcing reflect a proper description of the general concept city. The categories that were not correctly classified were left out since they did not refer to the correct meaning of the descriptions obtained by crowdsourcing and consequently could not reflect on the level quality of the description of city. These categories are not organized in a taxonomy and thus only the number of semantically equivalent categories with the Wikipedia resource was analyzed.

When removing duplicates in the categories from Word Reference for the hints, we obtained a total of 29 categories. From those, 18 (62%) directly corresponded to categories in the Wikipedia taxonomy for cities, modulo clear term alignment (like Food ≡ Cuisine). One other category corresponded to the Wikipedia taxonomy for regions. From the remaining 10 categories which did not have clear matches in Wikipedia, 8 were subconcepts of a category in the Wikipedia taxonomy (for example Mammal), while 2 were not present. For the taboos the total was 31 categories, from which 15 were in the city ontology, 6 on the other layers (region or country), 5 were subconcepts of categories in the city ontology, and 5 were not present. In both cases, 9 of the 12 first-level categories in the Wikipedia taxonomy were represented, either by themselves (in 6 cases) or by one of their subcategories.

Since we apply two different crowdsourcing techniques in this paper, it is also interesting to evaluate any differences in the resulting data sets of those techniques. From both datasets a total of 60 categories were obtained of which 60% are identical. Half of all non-corresponding categories for each data set represented specific concepts that would occur on a lower level of hierarchy and be subsumed by corresponding concepts, such as Eastern Religion is a subcategory of religion. The other half are categories that are very general, such as Sport, and are thus likely to occur on the highest level of hierarchy.

4.3.2 Ontology Alignment

The first evaluation step of Wikipedia was similar to the evaluation of the Word Reference categories in that we only considered semantically equivalent categories. For instance, 'meteorology' ' and 'climate' would be considered semantically equivalent categories, while 'snow' clearly is more specific. In addition, we also consider the level of hierarchy on which the categories co-occur and whether those correspond. This step serves to evaluate whether the semi-automatically built resources based on data from crowdsourcing approaches offers a similar level of detail as the manually created resource for describing the general concept of city.

The results of this evaluation are quantified in Table 6, which only includes correct hierarchy extractions. Thus, the number of deduplicated results was further reduced from the disambiguation step to 60 taboos and 48 hints with two correct levels of hierarchy. 'N' refers to no correspondence in Wikipedia, 'L0' to the most specific hierarchy level, 'L1' to the intermediate level, and 'L2' to the most general meaning of the Wikipedia city ontology.

	N	N (%)	L0	L1	L2
taboos	15	25%	11	19	15
hints	30	62%	7	7	4

Table 6: Comparing Resulting Ontology with Gold Standard Ontology from Wikipedia TOCs

When comparing the two different data sets, the seed ontology deriving from the mechanized labor-based data set, the taboos, shows a larger variety of types of categories that correspond to Wikipedia categories of cities with 45 in total. The categories and subclass relations obtained from the game-based crowdsourcing task correspond in 18 cases to Wikipedia elements on different levels of hierarchy. One reason for this lower coverage of hint categories in the gold standard ontology is the fact that they relate to more abstract Kyoto concepts. For instance, Kyoto#activity for cooperation is more general than categories that could be found in TOCs, such as Politics or Governance. Furthermore, WordNet Domains proved to show a high correlation with our gold standard ontology. Thus, the fact that more taboos directly correspond to WordNet domains is a second reason for the stronger correlation of the Wikipedia TOC and the Taboo seed ontology. It is interesting that the distribution of both types of ontologies is rather even across the three levels of hierarchy.

A direct comparison of the two data sets with each other, however, indicated a stronger variation of the results of the two crowdsourcing approaches. We compared the first level of hierarchy to each other, that is, the Kyoto concepts and WordNet domains either directly associated with the sense we obtained from the disambiguation techniques or associated with it on the next level of hierarchy. Deduplicating those concepts resulted in 46 concepts for the taboo words obtained from the mechanized labor-based approach and 31 concepts for the game-based approach. This comes as a little surprise since the first data set is larger than the second one. In total, 49% of the obtained 77 concepts are identical across both data sets on the first level of hierarchy. We classified the non-identical concepts into concepts and subconcepts. Concepts would likely be found on the most general level of hierarchy, while subconcepts would be found on a lower level, such as soccer as a subconcept of sport. The distribution of this classification is identical across the two data sets with 56% subconcepts and 44% concepts likely to be found on the highest level of hierarchy in a city ontology. Nevertheless, the type of subconcepts that can be found in the results obtained from the taboo dataset shows a slightly higher level of granularity. For instance, it contains concepts such as mountaineering that could only be found as sport in the hint data set.

5 DISCUSSION

The two distinct crowdsourcing techniques utilized in this approach both proved to be a valid and valuable source of input for the ontology engineering process. We found that the time needed to obtain data from the mechanized labor-based approach strongly exceeded the time for obtaining the same amount of data in a game-based approach. The former was running for more than a working week, while the latter achieved the same in just five sessions each a bit more than an hour. The incentive to participate in a game of Taboo seemed much higher. In fact, participants asked for the permission to play again after the first session, and four of the thirty participants joined a second session.

The nature of the game required the creation of Taboo words. When we investigated descriptions of cities online, we found very little useful data. Thus, we decided to crowdsource this first step. The fact that the results of this first method, the descriptions of the cities we called taboo words, are then used as input for the game-based approach is not ideal. While it reduces the overlap of the two data sets, it also creates an unwanted bias. Without this step the overlap of the data set might be much stronger and thus the resulting seed ontology much more similar than in this mutually exclusive way we propose. On the other hand, our major goal was not the comparison of the two techniques with each other. We were rather interested in the degree of overlapping concepts and hierarchical relations obtained from each data set and our manually created gold standard ontology from Wikipedia data.

The two approaches that we implemented for the extraction of senses from two different resources are accurate in that they have the correct sense for most of the words in the city descriptions. Word Reference is convenient because it already provides a classification of the senses in the form of a general domain label, however, there are many senses for which that classification is missing, which results in a great loss of useful data. A resource like this one but with a complete classification would be ideal for our purposes. For WordNet, the labeling feature is not immediately available, so more complex techniques need to be implemented to retrieve a classification of our data. In both cases there were many other senses available, so some kind of sense disambiguation is necessary.

The distributional approach returned impressively accurate results in some cases (for example, the pair hint-city *(star, Cannes)* matches with the definition *a prominent actor, singer, or the like, esp. one who plays the leading role in a performance*, which has the category *'Show Business'*). However, there are also some issues that should be resolved in future work. For example, using a simple comparison with the vector obtained from the average of the hint and the city causes that if any of the definitions includes the hint as a word, it will rank very high. Consequences of this are, for example, that *(go) jump in the lake, (used as an exclamation of dismissal or impatience)* ranks higher for *(lake, Lausanne)* than the actual sense of *boy of water*. This should be resolved by either exploring different ways of composing words in vectors, or by using a more complex combination, for example giving more weight to the city than to the hint when combining them.

The class-based approach is strongly biased by the static resources it utilized and thus, less attractive than the more dynamic distributional semantic approach. Moreover, the number of senses obtained from WordNet on only two levels of hierarchy can be very overwhelming. For highly ambiguous terms with several senses and upon querying YAGO, Kyoto, and the WordNet domain ontology, the number of retrieved categories and senses for an input word quickly reached more than 400. With the automated frequency-based and Word Reference category-weighted disambiguation we still obtained comparable results to the distributional approach. WordNet domains proved to be a highly reliable and disambiguating part of this approach as also found and suggested by [2]. First of all, their occurrence in the sense repository of a query shows that the queried word is very close to this high-level ontology associated with WordNet senses. Thus the queried word itself can be assumed to be more general in meaning than those in the same category that are not directly associated with a domain. Secondly, the domain proved to be highly accurate for the kind of disambiguation we needed. Finally, it associates the category of the description with exactly one further superordinate category and thus makes it more comparable to our Wikipedia ontology.

The strong difference of level of granularity in WordNet unfortunately propagates to the ontology concepts that are retrieved from Kyoto. For instance, DOLCE:endurant and Kyoto#organization are returned on the same level of hierarchy when querying the resource for the input word company. While for some approaches highly abstract concepts, such as DOLCE:endurant, are very useful for others, such as ours, they are too high-level. This also applies to the number of concepts and relations that are retrieved for each sense in WordNet. When considering two levels of hierarchy for an already disambiguated sense, the number of concepts can easily reach 14. Since a full evaluation of all relations and concepts retrieved goes beyond the scope of this paper, we decided to focus on more concrete levels of granularity, i.e., disregard upper level ontologies such as DOLCE for this approach, and only subclass relations. Kyoto returns a number of non-hierarchical relations, however, their evaluation goes beyond the scope of this paper. One alternative approach to handling this wealth of information might be a classification of concepts and relations based on machine learning, as implemented and proposed by [16]. Alternatively, crowdsourcing could also be applied to this step.

The comparison of the categories retrieved from Word Reference with the ones in the taxonomy built from Wikipedia shows that in most cases the labels match. Some of the ones that do not match directly are subcategories of Wikipedia labels, which seems to show that creating an organized taxonomy using the Word Reference taxonomies as seeds would be a promising direction. In other cases, the categories match with others in the Wikipedia taxonomies for *country* or for *region*, this should be taken into account when using this kind of approaches, since players tend to describe instances not only with their properties but also with properties from their parent categories.

A comparison of the two data sets resulted in a surprisingly high level of overlap of semantic categories given that they differ in size and the data of the first mechanized labor-based approach cannot be provided as data of the second approach since they represent the taboo words in the game. Thus, we concluded that the results of the two approaches are comparable, even if the taboo words lead to a higher level of granularity in the conceptualization of city descriptions. We found in the word sense disambiguation results that more specific descriptions provided more interesting hierarchies for the characterization of a city. For instance, *food* quickly became *substance* up the hierarchical classification ladder while *kiwi* mapped to *vine* and then *plant/flora* and *barbecue* to *nutriment* and then *food*.

6 CONCLUSION AND FUTURE WORK

This paper addresses ways to benefit from the diversity of people in the world by utilizing two distinct crowdsourcing techniques to gather data for ontology building. It further utilizes a third crowdsourcing platform, namely Wikipedia, to build an evaluation resource for the results obtained from the first two. The two word sense disambiguation methods used herein provide promising results for automating the step of concept formation and categorization of city descriptions. We also semi-automatically built a hierarchical backbone to the retrieved categories in order to facilitate their comparison with a manually created ontology for city descriptions based on the crowdsourcing platform Wikipedia. The results thereof show that the mechanized labor-based technique returns more specific categorizations and a more refined level of hierarchy. Nevertheless, the game-based approach returns very promising results and we believe that it is a more interesting way for the crowd to engage in a knowledge production tasks.

There are many directions of research that are derived naturally from this work. Some of the technical ones were pointed out in the previous section, while here we discuss more general questions that should be addressed.

First, the relation extraction part should be developed. Although there exist approaches that tackle this problem in particular, both with automatic and crowdsourcing techniques, their adequacy to our problem should be analyzed, since they are not particularly designed to identify relations between a concept and its attributes. The classification part, for which we provide automated methods here, could also be crowdsourced.

Second, the choice of using an implicit crowdsourcing method could be justified empirically by comparing it with an explicit technique for the same task, something that we kept for future experiments for now. To this end, we should perform a third experiment in which users are asked directly to name properties of the general concept city.

Finally, the true diversity of the obtained domain knowledge could be further explored by building clusters based on common traits of participants, such as country of origin or age, and comparing the results of individual clusters to each other. This also provides a large number of individualized domain ontologies that are highly comparable and might provide some insights into the diversity of knowledge production. Furthermore, conducting comparable experiments with non-English speaking crowds and comparing the results obtained from multilingual corpora obtained from crowdsourcing could be an interesting direction for further research.

ACKNOWLEDGEMENTS

This research has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 /ESSENCE: Evolution of Shared Semantics in Computational Environments/.

References

[1] P. Basile, A. Caputo, G. Semeraro, and F. Narducci, 'Uniba: Exploiting a distributional semantic model for disambiguating and linking entities in tweets', *CEUR Workshop Proceedings*, **1395**, 62–63, (2015).

[2] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta, 'Revising the wordnet domains hierarchy: semantics, coverage and balancing', in *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pp. 101–108. Association for Computational Linguistics, (2004).

[3] Hafedh Chourabi, Taewoo Nam, Shawn Walker, J Ramon Gil-Garcia, Sehl Mellouli, Karine Nahon, Theresa A Pardo, and Hans Jochen Scholl, 'Understanding smart cities: An integrative framework', in *Proceedings of the 45th Hawaii International Conference on System Science (HICSS)*, pp. 2289–2297. IEEE, (2012).

[4] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei, 'Leveraging the Wisdom of the Crowd for Fine-Grained Recognition', *Ieee Transactions on Pattern Analysis and Machine Intelligence*, **38**(4), 666–676, (April 2016).

[5] Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, and Anthony Levas, "'dr. detective": Combining gamication techniques and crowdsourcing to create a gold standard in medical text', in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, pp. 16–31, (2013).

[6] Kai Eckert, Mathias Niepert, Christof Niemann, Cameron Buckner, Colin Allen, and Heiner Stuckenschmidt, 'Crowdsourcing the assembly of concept hierarchies', in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pp. 139–148, (2010).

[7] Florian Hanika, Gerhard Wohlgenannt, and Marta Sabou, *The uComp Protégé Plugin: Crowdsourcing Enabled Ontology Engineering*, 181–196, Springer International Publishing, 2014.

[8] Rubén Izquierdo Beviá, Armando Suárez Cueto, German Rigau Claramunt, et al., 'Word vs. class-based word sense disambiguation', *Journal of Artificial Intelligence Research*, **54**, 83–122, (2015).

[9] Lili Jiang, Yafang Wang, Johannes Hoffart, and Gerhard Weikum, 'Crowdsourced entity markup', in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web-Volume 1030*, pp. 59–68, (2013).

[10] Vladimir I Levenshtein, 'Binary codes capable of correcting deletions, insertions, and reversals', in *Soviet physics doklady*, volume 10, pp. 707–710, (1966).

[11] Miguel Angel Luengo-Oroz, 'Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears', *Journal of Medical Internet Research*, **14**(6), e167, (2012).

[12] Alexander Maedche, *Ontology learning for the semantic web*, volume 665, Springer Science & Business Media, 2012.

[13] Emir Muñoz, Aidan Hogan, and Alessandra Mileo, 'Triplifying wikipedia's tables', in *Proceedings of the First International Conference on Linked Data for Information Extraction - Volume 1057*, LD4IE'13, pp. 26–37, (2013).

[14] P. Nasirifard, S. Grzonkowski, and V. Peristeras, 'Ontopair: Towards a collaborative game for building owl-based ontologies', volume 351, pp. 94–108, (2008).

[15] Natalya F Noy, Jonathan Mortensen, Mark A Musen, and Paul R Alexander, 'Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow', in *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 262–271. ACM, (2013).

[16] Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader, and Michael Schroeder, 'Formalizing biomedical concepts from textual definitions', *Journal of biomedical semantics*, **6**(1), 1, (2015).

[17] Cristina Sarasua, Elena Simperl, and Natalya F. Noy, *CrowdMap: Crowdsourcing Ontology Alignment with Microtasks*, 525–541, Springer Berlin Heidelberg, 2012.

[18] Neil Savage, 'Gaining Wisdom from Crowds', *Communications of the Acm*, **55**(3), 13–15, (March 2012).

[19] Katharina Siorpaes and Martin Hepp, 'Ontogame: Towards overcoming the incentive bottleneck in ontology building', in *Proceedings of the 2007 OTM Confederated International Conference on On the Move to Meaningful Internet Systems - Volume Part II*, OTM'07, pp. 1222–1232. Springer-Verlag, (2007).

[20] Stefan Thaler, Elena Simperl, and Katharina Siorpaes, 'SpotTheLink: A game for ontology alignment', in *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, volume P-182, pp. 246–253, (2011).

[21] Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli, 'Validating and extending semantic knowledge bases using video games with a purpose.', in *ACL (1)*, pp. 1294–1304, (2014).

[22] Johanna Völker, Daniel Fleischhacker, and Heiner Stuckenschmidt, 'Automatic acquisition of class disjointness', *Web Semantics: Science, Services and Agents on the World Wide Web*, **35**, 124–139, (2015).

[23] Luis von Ahn, 'Duolingo: learn a language for free while helping to translate the web', in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 1–2. ACM, (2013).

[24] Luis Von Ahn and Laura Dabbish, 'Designing games with a purpose', *Communications of the ACM*, **51**(8), 58–67, (2008).

[25] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, 'recaptcha: Human-based character recognition via web security measures', *Science*, **321**(5895), 1465–1468, (2008).

[26] Wilson Wong, Wei Liu, and Mohammed Bennamoun, 'Ontology learning from text: A look back and into the future', *ACM Computing Surveys (CSUR)*, **44**(4), 20, (2012).

[27] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung, 'A survey of crowdsourcing systems', in *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*

and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 766–773, (2011).

[28] Maayan Zhitomirsky-Geffet, Eden S Erez, and Bar-Ilan Judit, 'Toward multiviewpoint ontology construction by collaboration of non-experts and crowdsourcing: The case of the effect of diet on health', *Journal of the Association for Information Science and Technology*, (2016). Online version.

[29] James Y Zou, Kamalika Chaudhuri, and Adam Tauman Kalai, 'Crowdsourcing Feature Discovery via Adaptively Chosen Comparisons', in *Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP) 2015*, pp. 198–212, (2015).