Towards Building Ontologies from Crowdsourced Data

Paula Chocron[1] and **Dagmar Gromann**[2]

Abstract. Building computational conceptual models that are flexible and context-independent is an important challenge given the growing interest in cross-domain applications that need to deal with real-world objects. While crowdsourcing methods have been used extensively in ontology engineering and evaluation, few approaches apply these methods to retrieve data. However, retrieving data from a large crowd promises to elicit the whole spectrum of associative knowledge humans use to cognitively represent concepts. In this paper, we propose to use two crowdsourcing techniques - a mechanized labour-based and a game-based approach - as a data acquisition method and a semi-automated approach for building a domain-specific ontology from this data for the concept *city*. Our approaches are indirect, asking participants to describe instances of cities. In a second phase, we implement techniques to extract semantic classes from the obtained data. We evaluate our results against another collaboratively built ontology for *city*, extracted from Wikipedia. We compare the techniques used, analysing benefits and drawbacks for each one.

1 Introduction

The past decade has witnessed an increasing interest in integrated visions that propose a tight and optimised relation between Web technologies and everyday real-life domains, such as the Internet of Things [1] or Smart Cities [7]. These efforts require sound computational models that represent concepts ubiquitous in our everyday life. Moreover, since different domains interact in these approaches, the models need to be as context-independent as possible, very flexible, and easy to adapt to diverse applications or to change and evolution.

The computational modeling of these concepts is a difficult problem that has been largely discussed from both the philosophical and the engineering perspective. Concepts that people can easily understand in informal conversations can be challenging to define and model formally, something that seems to be particularly true for spatial and geographic concepts, as it can be seen in the well-known discussion about the definition of *forest* (see, for example, [4]). An alternative solution would be relying on a group of *experts* to build the concept descriptions. This is, on the one hand, a time- and cost-intensive task [31]. On the other hand, it is not free of arbitrariness, since the resulting conceptualisations can be biased by the view of the person(s) in charge of building them. We believe that relying on a large population for acquiring associative knowledge drastically reduces this potential bias.

In the past years, *crowdsourcing* techniques received increasing attention as methods that can overcome these two problems. Crowdsourcing consists of 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 automate 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.

In this paper we propose a method that uses two crowdsourcing techniques – a mechanized labour-based and a game-based approach – to build concepts for ontologies, identifying different types of knowledge that humans use to describe instances. In line with results of Parasca et al. [22] we observe that people extensively rely on prior knowledge as well as synonymy, antonymy, and hypernymy to describe concepts. This kind of task can benefit from collaboration, as it is discussed in [15], where the authors compare conceptual maps created individually and collaboratively, concluding that the latter ones are of higher quality.

Using crowdsourcing can provide a description that reflects the collective perception of a concept, identifying categories that are socially relevant, but not immediate when explicitly trying to define them. For instance, five participants in the mechanized labour-based approach used *café* to describe *Paris*. In the approach using a word guessing game, four participants used *baguettes* to describe *Paris*, which was guessed correctly upon this first hint in four different games. When asked directly to describe *city* it is unlikely that participants would provide similarly specific associations. Our approach is still preliminary, but the data obtained[3] can be used as a pattern to describe concepts in applications where the social aspect is relevant or as a human-created standard to evaluate automated techniques. We also believe that the game-based approach can provide a valuable method for cognitive applications, since it can specify the type of knowledge to be elicited in a way that is entertaining to participants.

We chose to perform our experiments using the concept of *city*. The concept of *cityness* has been extensively discussed in the urbanistics literature, remarking its social and dynamic

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

[3] Data obtained from and source code used in this approach can be found at `https://github.com/paulachocron/CrowsourcedKnowledgeAcquisition`

aspects. In particular, we consider *city* to be a 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.

After introducing related work, Section 3 explains the design and implementation of two crowdsourcing methods that retrieve data by asking participants to describe instances of a city. The mechanized labour-based technique asks participants directly, while the second one presents the task in the form of a game, making it more attractive to participants. This game-based method can be seen as an extension of the first one, that could potentially complete the descriptions obtained. In Section 4 we explain a post-(crowdsourcing)-processing phase, in which we implement two methods to automatically extract categories related to cities from the crowdsourced data. We evaluated our approach by comparing its results to another crowdsourced description of cities that we extract from the Wikipedia Tables of Contents of city pages.

2 Related Work

Crafting ontologies manually is a costly task, and the obtained results are not free of arbitrariness. For these reasons, the field of ontology learning has been extensively studied in the past years [18]. Many of these approaches, particularly those from the first years of the area's development, rely on predefined patterns and rules or static resources, such as WordNet [31]. 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 [23] or to develop axioms extracted from text [27].

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. [5] associate senses with a WordNet domain ontology they create and which we also use herein to classify words. A similar idea is presented by Izquierdo et al. [14] 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. Alternatively, distributional semantic approaches have been investigated for word sense disambiguation with context-poor data sets. For instance, Basile et al. [3] 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 [3] the authors use addition.

The use of crowdsourcing techniques in ontology learning has received considerable attention in the past few years. One prominent example can be found in the approach of Hanika et al. [13], who present approaches that solve some specific tasks in ontology engineering via crowdsourcing, mainly related to specifying relations between terms and verification. Eckert et al. [11] propose a method to crowdsource concept hierarchies, asking users questions about how concepts are related. In general, these methods start from an existing set of concepts and attempt to add relations collaboratively, without allowing the discovery of new concepts. Crowdsourcing has also been applied to ontology evaluation (both for the subclass-superclass hierarchy [21] and for entire ontology statements [32]) and alignment [24].

In an effort to make crowdsourcing tasks more appealing to participants, the idea of *gamification* was introduced that presents the problems to be solved in form of an interactive game, which is thought to foster motivation to participate [25, 29]. Some well-known examples are Duolingo [28], an approach to crowdsource the translation of the Web, and reCAPTCHA [30], a method for digitizing paper copies of documents. Parasca et al. [22] utilize a guessing game to elicit associative knowledge to define words provided to the players. They analyse the type of associative knowledge obtained this way and suggest their data set to be used to evaluate distributional approaches, which makes their method more focused on linguistics. Verbosity [29], a game to elicit commonsense facts in a structured way, restricts the elicitation of knowledge to specifically related items in order to obtain already related expressions. We decided against this design to have a non-restrictive elicitation of associative knowledge.

Individual ontology engineering tasks have been crowdsourced as games as well, such as for classification and population [26]. In [20] 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 Crowdsourcing Concept Descriptions

We propose two different crowdsourcing mechanisms to collect descriptions of city instances that will later be used to build the concepts related to city. The approach, from the data collection to the evaluation, is illustrated in Figure 1, where rectangles are steps and circles are the different techniques that we explore.

Crowdsourcing methods can be divided into two major categories [10]. *Explicit* methods are those in which participants are asked directly to perform some task, and then their work is combined to solve a larger problem. In the *implicit* approach users are given a task that is only indirectly related to the problem to solve; then some kind of post-processing is applied to extract the desired information from the retrieved data. The implicit approach allows designers to make the task more attractive for participants, for example by presenting it as a game. For some tasks, implicit approaches can result in more interesting and fine-grained results, and for some experiments it is crucial that participants do not know the final objective to avoid cognitive biases.

We used two different implicit crowdsourcing methods, in which we ask participants to describe specific instances of cities either as a direct question or as part of a game to obtain a general characterization of *city* as a concept. We believe that using an implicit approach for our problem can lead to richer and more fine-grained ontologies than the explicit one. However, the direct comparison to this kind of task is yet

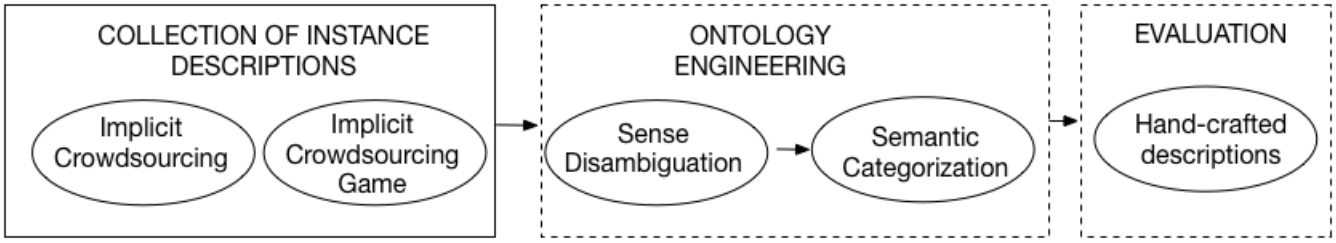

Figure 1. Overview of Concept Formation Method

to follow. The kind of technique we propose here is particularly applicable when describing abstract concepts that do not have a clear physical correspondence, where the properties are less evident. To form concepts, we group the descriptions we obtain from the crowdsourced data by their sense that we disambiguate and by their semantic category that we extract.

A first approach, which we call *mechanized labour-based knowledge acquisition*, uses a popular crowdsourcing online platform (*CrowdFlower*) in which users can register and complete mechanical tasks in exchange for payment. The task we proposed consisted in directly asking participants to provide the first ten words they associate with a city name displayed to them. In the second approach we adapted the well-known Taboo game, in which one player describes a city instance to the other player without using any of the words on a list that is provided to them. This second technique was developed as an extension to the first one, and it therefore uses the results of the first one as words that cannot be used in the game in an effort to obtain more fine-grained, original descriptions of *city*.

For both methods we used as input a list of city names that we built with input from online listings of popular cities retrieved by a search engine query. Participants had to be first language English speakers since the data we wished to obtain are in English. They were always given the possibility to skip a city if they were not familiar enough with it. They were explicitly instructed to only use common nouns with adjectival/verb modifiers. 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. Participants were explicitly instructed to comply with this input restriction, in the mechanized labor-based approach even tested on them, and non-conforming characterizations were omitted from the final data set. In the following subsections we explain each of the two approaches in detail.

3.1 Mechanized Labour-Based Knowledge Acquisition

Description

In the first of our approaches we ask participants directly to provide descriptions of a given city. This was implemented using the online crowdsourcing platform *CrowdFlower*. In a crowdsourcing task, participants were provided with the name of a city (for example, Paris), its country name (France), and its latitude and longitude (48.85° N, 2.35° E), plus ten input fields for city descriptions as illustrated on the right hand side of Figure 2. The instructions clearly described the task and specified the input restrictions to describing cities by using common nouns with possible adjectival/verbal modifier as description, no words other than English including loan words, and a categorical omission of personal opinions. For instance, *stinky cheese* to describe *Paris* by means of one of its most prominent food exports would be permissible in the sense of providing a common noun in English but clearly violates the instruction of not providing a personal opinion. Each page in the task displayed a total of five city instances to each participant, asking for their descriptions following the specified input restrictions.

The left hand side of Figure 2 shows the first part of a page displayed to the participants, where they could indicate whether they had heard of the city, had been there, or are not familiar with it. The right hand side of Figure 2 shows the view when clicking one of the first two options, where the user is prompted to provide a maximum of ten words they associate with that city. We uploaded a list of 300 cities that we collected from online resources and there was no limit to how many descriptions participants could provide.

We took several measures to ensure the quality of the results for this crowdsourcing technique. First, we conducted a pilot study to evaluate the kind of results we were to expect and to improve our test setup. Second, we asked each participant 20 test questions to ensure their ability to comply with the instructions regarding the input restrictions. For example, we asked participants whether Breaking Bad is an adequate description of Albuquerque, USA, to test the ability to differentiate common from proper nouns. Only participants with an accuracy above 70% on the test questions were permitted to complete city descriptions. Finally, answers on which participants spent less than ten seconds on the question were not considered, a setting enabled by the platform.

This measure was taken because it can be reasonably assumed that a participant who spends less than on average ten seconds on the description of five city instances has not taken the time to provide enough descriptions or delivers poor quality. For instance, if a participant provides *cathedral* for all five cities and no other input it might be possible to do so within ten seconds, but this does not mean that *cathedral* is an adequate description for those cities. It is highly unlikely that a human being is able to process five city names and type at least one association for them within ten seconds. This is a common quality measure in crowdsourcing approaches and in combination with the test question turned out to be a very effective way to obtain high quality contributions. The pilot study was particularly important since the improved actual study received a better evaluation by the participants, which increased participation. The more participants, the faster the

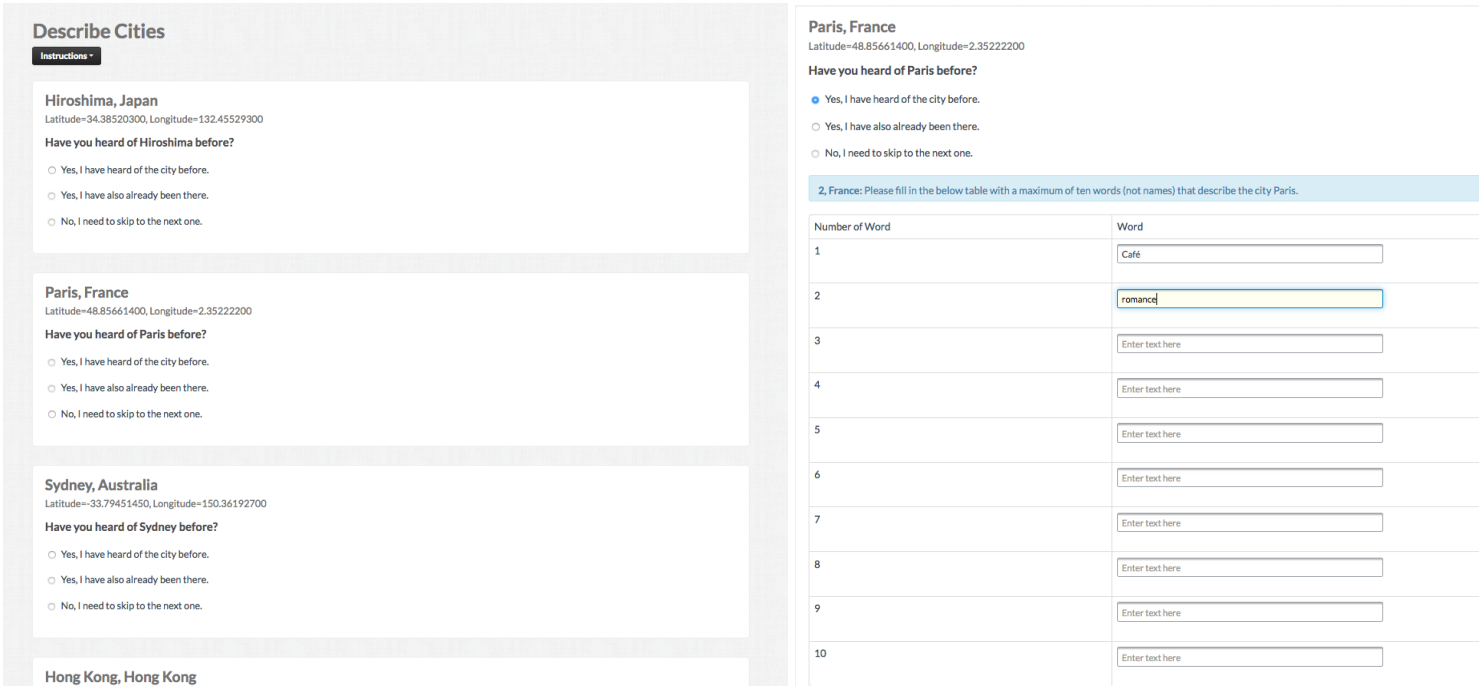

Figure 2. Example Question on *CrowdFlower*

specified number of targeted descriptions can be achieved.

Results

The task was online on *CrowdFlower* for six days asking participants to describe a total of 300 cities. At the end of the task, 82 participants mainly located in the United States (65%) and the United Kingdom (30%) curated a total of 3,616 descriptions for 275 of the 300 cities. For the remaining 25 cities, no descriptions were provided since the participants had the option to indicate that they did not know a city displayed to them in the task. We did not limit the number of descriptions that could be provided by an individual participant in the overall task and some top contributors provided up to 85 descriptions. The proportion of participants that passed all the trust tests (70% accuracy on the test questions and answer time not too short) was 91%, which shows that the task was well designed and accessible for participants. This argument is further supported by a final contributor satisfaction rating of 4.2 out of 5 potential points for our task.

The 3,616 descriptions we obtained for the 275 cities contained duplicates. In a first processing phase, we kept duplicate descriptions across the data set but de-duplicated the ones for each city. For instance, we would de-duplicate the five mentions of *café* relating to *Paris, France* to one while keeping *café* as a description of *Tangier, Morocco*. This was done automatically by applying similarity measures from the WordNet Similarity for Java (WS4J) library[4] combined with the Levenshtein distance [17]. This approach was also used to identify the descriptions that were provided most frequently, for example, there are in total 576 descriptions that were provided by more than one user. For instance, for *Paris* the description *café* was identified as one of the most frequent ones as it was provided by five distinct participants. To reduce the bias of individual associations regarding specific city instances, we decided to choose only those frequent descriptions for the concept formation task.

3.2 Game-Based Knowledge Acquisition

Our second crowdsourcing technique presented the city description task as a game that was designed by adapting the popular board game Taboo™[5], which we call GUESSence. First, we will briefly describe the original Taboo™ game, then our version, and finally the results we obtained from the game.

Description of the Taboo Game

The popular board game Taboo is a word guessing game, where an even number of players is grouped into competing teams. One player of one team draws a card and describes a word it provides, with the objective of having their team colleagues guess the word without using the word itself or five related words indicated on the same card. The five related words are called taboo words and neither the words themselves, their components, nor their inflectional variants can be used. For instance, if the word is *basketball*, it is not permitted to use *ball*, *baskets*, or *basketball* to describe it, and neither is it permitted to use any of the given taboo words, for instance *court, player, bounce, baseline, game*. The goal is to have your team guess as many words as possible within the alloted times. Players take turns at describing words and after the alloted time is over, it is the turn of the next team. Each team receives one point for a correct guess and the team with most points at the end of the game wins.

[4] `https://code.google.com/archive/p/ws4j/`

[5] `http://www.hasbro.com/common/instruct/Taboo(2000).PDF`

Description of GUESSence

In our two-player version of 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. As a further restriction, the describer is not allowed to use any of the phrases that are provided as *taboo words* along with the city name and the country it is located in. The objective of the game is that the guesser finds out which city is being described. The game is collaborative, since when the guesser names the correct city both players win. As an example depicted in Figure 3, if the city to describe was *Paris, France*, the describer would get the list of taboo words *capital, café, romance, croissant, art, tower, fashion, museum, palace, terrorist, cathedral*. Then a game could be developed as exemplified in the following, where the describer starts the game by saying *river* and the guesser provides the first guess. When the guesser types the correct city name the game was successful and both players are automatically assigned to a new game.

describer:	River
guesser:	London
describer:	Famous pastries
guesser:	Vienna
describer:	Hunchback
guesser:	Paris

The taboo words for this game were the descriptions obtained from the first data collection method, ensuring that there was no overlap between the data set gathered with the first knowledge production method and this one. Additionally, we intended to analyse if this would trigger the use of descriptions belonging to different categories in an effort to find original descriptions. We only used cities for which we had enough taboo words, omitting all cities with less than five descriptions in the first task. This reduced the number of 275 cities from the first task to 244 in the game. In an effort to make each game equally challenging, we decided to include some hand-picked additional salient descriptions as taboo words if they were missing, such as *canal* for *Venice*. By hand-picked we mean we complemented the most frequent descriptions of the first task by associations that only one or two *CrowdFlower* participants had who indicated they had visited the city before.

We determined the number of taboo words on the basis of the number of descriptions provided for each city by the crowd of the mechanized labour-based approach. This was based on the assumption that the more associations participants of the first crowdsourcing method had with a specific instance of a city, the easier it was to retrieve those associations and the more well-known specific associations for that city might be. To make the description of each city instance equally challenging in the game-based crowdsourcing approach, we kept more taboo words for cities with a large number of associations while keeping fewer taboo words for cities with a lower number of associations from the first crowd. Concretely, cities with more than 25 descriptions were equipped with 12 taboo words in the game, cities with 20 to 25 descriptions were assigned 10 taboo words, and 8 taboo words were provided for cities with less than 20 descriptions from the first crowd on *CrowdFlower*. Cities with less than five descriptions from less than two participants were omitted in the game-based approach since this can be seen as an indication that not enough associations could be elicited from the crowd and thus the city might be more challenging to describe in the game-based approach than the other city instances.

Crowdsourcing

Since the game requires collaboration, participants had to be online simultaneously, which made the use of crowdsourcing platforms challenging. As we observed in the first data collection method, the number of participants that simultaneously accessed the crowdsourcing task rarely exceeded two at a time. Thus, we had to devise an alternative method to recruit participants. We contacted colleagues within the ESSENCE project[6] at the School of Informatics at the University of Edinburgh, and asked them to send out an invitation to participate on internal mailing lists and invite personal contacts. This local recruitment should ensure that participants in the game were first language English speakers in line with the first crowdsourcing method and this was the only partner institution located in an English speaking country. As an incentive to participate, we offered a small shopping voucher. To avoid personalized hints when participants knew each other, the method for joining two players in a game was automated and both players had no means of identifying who the other player was.

We developed an online platform[7] and pre-scheduled game sessions with up to nine players at a time, who were assigned automatically and anonymously to two-player games. 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 view that is illustrated in Figure 3. The game starts with the describer providing a hint and ends with the correct guess from the guesser. Players were newly assigned automatically and anonymously to each game. The same input restrictions as in the first crowdsourcing method were applied for the hints of the describer, which was additionally enforced by automated warning messages. For those warnings, we implemented several named entity recognition processes using NLTK [6] and the Stanford Named Entity Recognizer (NER) [12]. Upon a correct guess, the interface automatically redirects both players to a success page. Should the automatic detection of a correct guess fail because of spelling errors or other reasons, we provide a "Guess Correct" button on the interface. If a player decided to leave the game and start a new game, the other player was automatically informed by the system that the game was over.

A total of 30 participants played the game in five online sessions. This resulted in 316 games, of which 174 were successful, i.e., the city was guessed correctly. We decided to limit our data set for the concept formation process to successful games only. 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. The lengths of the games varied substantially. At times players were able to

[6] `https://www.essence-network.com/`
[7] `http://taboo.iiia.csic.es`

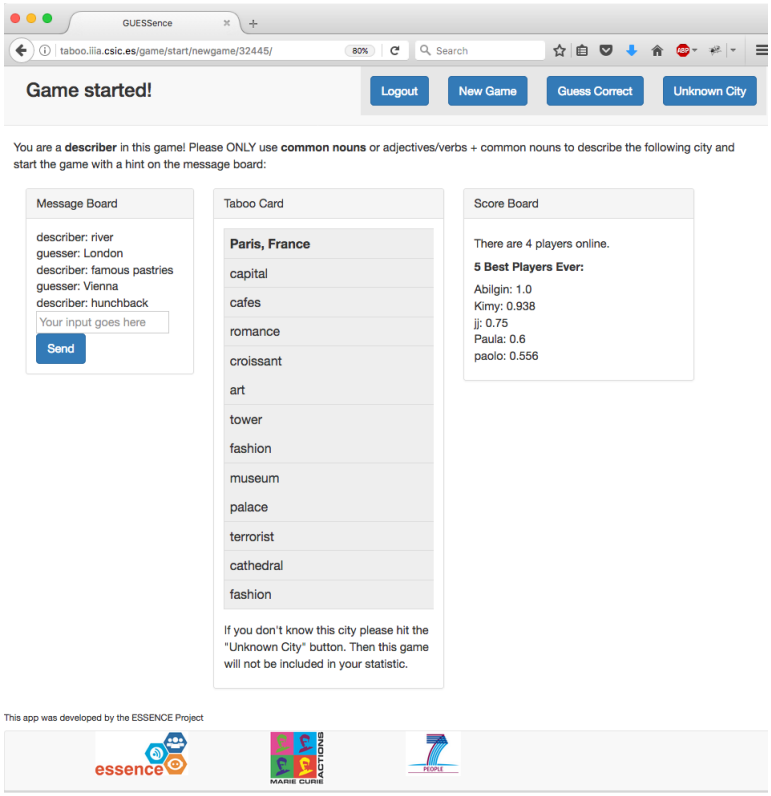

Figure 3. Example GUESSence Game: Describer View

guess the correct city name upon receiving the first hint, such as *baguettes* immediately triggered the response *Paris* when indeed *Paris* was the target city. In other games players required up to ten hints to correctly guess the target city. It can be presumed that first hints that immediately and in several games trigger the correct target city are more associated with the instance then hints that require further descriptions to lead to a correct result. For instance, *baguettes* as a first hint for *Paris* lead to four successful games with different players, which means that *baguettes* is closely related to the city. We could observe a strategy of guessers to first provide the capital of a country they presumed the target city to be in, so it could be argued that *baguettes* is more related to the country itself then the city. Indeed, the types of knowledge used by the participants to describe a city came from different categories, such as regions, continents, climate, food, fauna, flora, among many others.

Those successful games were manually evaluated by 12 ontology engineers and researchers regarding their compliance to the restriction to common nouns and the rules of the game, e.g. not containing a Taboo word, which reduced our data set of successful games 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 concept formation method.

4 Building Ontologies

Building ontologies from natural language descriptions is a four step process that requires *concept formation* (e.g. [8]), extraction of hierarchical relations (e.g. [9]), learning non-hierarchical relations (e.g. [23], and finally extracting axioms (e.g. [27]). In this paper we focus on the first step of forming concepts from the instance descriptions we obtained in the two crowdsourcing processes. Concept formation is usually understood as the process of grouping terms into classes based on their shared semantic properties [31]. This involves detecting term variants, which is commonly done by using predefined background knowledge [31].

In our data, the available semantic property is the relation of each word to a city instance. Thus, we need to disambiguate each word and then group them by means of their related senses that we retrieved form existing lexical resources. We also retrieved predefined semantic categories into which data from both crowdsourcing approaches are classified. In this section we will differentiate between the data from the two approaches by referring to data from the mechanized labour-based approach as *taboos* and descriptions from the game-based approach as *hints*. To continue with our example, the hint *baguette* is related to *Paris* and the sense we want to obtain is that of "a long, narrow loaf of French bread" with the semantic category *Food*, where the category provides the general characteristic people use to describe a city. This is the ultimate objective of our approach: find the types of concepts that people associate with the general concept *city*. To evaluate our obtained concepts, we use an ontology that we created manually based on Wikipeda Tables of Contents (TOCs) of specific instances of cities. We will first explain how we obtained this evaluation ontology and then detail the approach to disambiguate and classify the crowdsourced data.

4.1 Building an Evaluation Ontology

When searching for characteristics of specific cities, we found that the Table of Contents (TOCs)[8] of Wikipedia pages for city instances shares some interesting characteristics with the kind of ontology that we attempt to build. ToCs are organised hierarchically, including subsections and even lower levels. As a collaborative encyclopedic resource where each page is quality assured by several community members, Wikipedia could be considered a long-term crowdsourcing approach and due to its quality assurance can be considered an excellent testbed for the concepts we form. Thus, we decided to manually build an ontology from TOCs of Wikipedia pages as a gold standard ontology for the characterization of city.

To build this ontology, four ontology engineers extracted Wikipedia TOCs of 20 randomly selected cities from our list of 300 cities. We merged the TOCs of those city 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"). In general it was easy to achieve an agreement, which suggests a high degree of consistency in Wikipedia's ToCs. Since many participants in the crowdsourcing processes used descriptions related to the country or region the city is in to describe the city, we repeated the ontology construction process from TOCs for countries, regions, and continents, which provided us with a four-layer gold standard ontology we used to evaluate our ontology built from crowdsourced data[9].

4.2 Concept Formation

To group our data into semantic categories, we retrieved the available senses and classifications for each noun and noun phrase from WordNet and an online dictionary, called Word Reference[10]. We opted for this combination since we found that semantic categories provided by Word Reference frequently complement WordNet domains [5]. An unexpected level of complexity was found at this point, triggered by the multiplicity of senses that may exist for each hint. In many cases, different senses corresponded to different categories; to continue with our example, baguette can be interpreted as "a small convex molding especially one of semicircular", which is classified as Architecture instead of the desired category Food for this hint. To overcome this problem, we implemented and compared several word sense disambiguation techniques. Since the focus of this paper is on crowdsourcing and ontology building, we will only briefly report on the most successful technique we applied.

By means of a word sense disambiguation approach based on distributional semantics adapted from [2], we identified the meaning that is semantically closest to the the general concept of city from WordNet and Word Reference definitions. We used each word of our input data in a vector representation and composed it with the vector of the city name. In a second step, we created a vector representation of each extracted definition and compared it with the first vector. The one combination of a definition vector with the $(word, city)$ that returned the highest similarity measure was the sense we chose automatically to be closest to the general concept of city. For instance, in the case of the two *baguette* senses related to *Food* and *Architecture*, the disambiguation process compares the vector for $(baguette, Paris)$ with the vector for each sense that is composed from the words in their glosses and returns a higher similarity for *Food*. Our composition method for vectors is that of averaging as implemented in the word2vec library [19]. All obtained senses were manually evaluated by two raters and only senses that both raters agreed upon were kept. We determined an F-Measure of 82% (164 taboos from 199) for the first crowdsourcing platform data set and 80% (89 hints from 112) for the second game-based data set. 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'.

We implemented two different approaches to extract predefined categories from descriptions of particular cities: categories from an online dictionary and ontology classes mapped to WordNet senses. In the first technique, we exploited the fact that Word Reference associates words with general labels that can be generally seen as its superordinate class. For example, Sushi is labeled as Food as is baguette which groups them together. To use this information, we first extracted all nouns in the crowdsourced city descriptions and then retrieved all existing glosses and categories from Word Reference where available. We used our distributional semantic disambiguation approach on the glosses to which labels were assigned to find the best gloss and extract its associated category. Our second approach consists of using WordNet to obtain the semantic classes mapped to its senses. Due to the fine-granular nature of WordNet senses, it was necessary to use ontologies associated with WordNet synsets to obtain general categories, as it is described in detail below.

In general, both Word Reference (WR) and WordNet (WN) contained definitions for the words in the descriptions, although there were some exceptions, most of which corresponded to words that can be considered from foreign languages (for example, names of foods or sports). For the majority of the words that were defined, the resources included the sense that was intended by the describer as one of the possible definitions, as it is shown in Table 1. We use the name *hints* to refer to data obtained in the game from the describer and *taboos* for the results of the first mechanized labour-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 the utilized resources. This is the case of, for example, using sack to describe Sacramento. The evaluation was done manually by three researchers.

We also measured the number of available correct categories in Word Reference as depicted in Table 2. In this case the values are 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

[8] https://en.wikipedia.org/wiki/Help:Section#Table_of_contents_.28TOC.29

[9] The ontology is available at `https://github.com/paulachocron/CrowsourcedKnowledgeAcquisition`

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

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

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

and the accuracy we obtained exceeds 80% for both data sets. This means that 80% of the extracted semantic categories were correctly assigned, which is a good result given the lack of context.

	Available	Not Available	Correct	Accuracy
taboos	132	64	119	0.90
hints	77	42	65	0.84

Table 2. Availability of categories in Word Reference

We also evaluated the Word Reference categories by comparing them with the city ontology from TOCs. We performed this evaluation only for 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. These categories are not organized in a taxonomy, and thus only the number of semantically equivalent categories with the Wikipedia resource was analysed.

When removing duplicates in the categories from Word Reference, we obtained a total of 29 categories for the hints and 31 for the taboos. In Table 3, we show the matching proportion with the Wikipedia taxonomy of cities, modulo obvious term alignment (such as Food ≡ Cuisine). We count in N the categories that did not match, and then present the ones that were present directly, the ones that were subconcepts of a present one (for example Mammal) and the ones that were present in the Wikipedia taxonomies for region or country. 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.

	N	N (%)	Present	Subconcepts	Other tables
taboos	5	16%	15	5	6
hints	2	6%	18	8	1

Table 3. Evaluation with Gold Standard Ontology for the WR categories

To classify the WordNet definitions, we used existing mappings to ontology concepts in YAGO[11] and the Kyoto mapping to WordNet [16] to extract classes associated with descriptions. Using the disambiguated WordNet senses we first extracted all WordNet Domains (WNDs) [5] associated with them from YAGO. This resulted in 17 for the hint data set and 6 for the taboo data set, of which one for each data set was incorrect. Those numbers already refer to de-duplicated senses used as input, that is, each WND was only counted once. Where there was no WordNet domain, we queried associated Kyoto classes and used the Word Reference categories to automatically select the best class. This was then again evaluated manually. We extracted a total of 78 classes for senses for the hint and taboo data set with an accuracy of 80% for the first and 91% for the second.

In a nutshell, our approach to concept formation consists of disambiguating individual specifications of city instances and grouping them by semantic categories. We obtain the semantic categories from WordNet domains and Word Reference categories. Those are then compared to concepts of the Wikipedia ontology we created as a means to evaluate our approach. Each concept created by this method comprises numerous characterizations of city instances. The set of concepts we obtain characterize the general concept *city*.

4.3 Comparison of Acquisition Approaches

Since the data of the mechanized labor-based acquisition method served as input to the game-based approach, the data sets we obtained are mutually exclusive. On a higher level of abstraction, that is, on the level of the semantic categories we obtained from Word Reference, the overlap between the obtained categories was 60%. Many of the remaining 40% that did not correspond across the two data sets were more specific concepts. For instance, the hints resulted in *Religion* while the taboos also returned *Eastern Religion*. Another example of the reverse phenomenon is that the hints only lead to *Clothing* while the taboos also lead to *Textiles*. Based on some concepts returned, such as *Drugs*, and an examination of the crowdsourced data we found that the game-based approach tends to provide more informal descriptions than the mechanized labour-based approach. This could be explained by the fact that the latter task itself has a higher degree of formality where participants are formally paid, while a word guessing game can be considered an informal situation. However, in some cases the use of informal words to avoid using the forbidden taboo words has been observed as a strategy. For instance, a player used *haggling* to describe *Bangkok* since the crowd on the online platform had provided *bargaining* as a taboo word.

The paid crowdsourcing task was conducted mainly by participants in the United States and also several participants from the United Kingdom (UK). In contrast, the game-based method was exclusively based on participants from Edingburg, UK. It could be expected that this difference in country of residence of the participants would affect the nature of the data in the sense of introducing a cultural bias. However, with very few exceptions no local expressions or dialect could be observed in the data. Interestingly the difference between American and British English was used to describe specific cities, such as an American participant using *soccer* to describe *Newcastle* in the UK. The same phenomenon occurred with other languages, such as a British participant in the labor-based approach who used the German *s-bahn* (rapid transit railway) to characterize Berlin in Germany or a player in the game describing Osaka with *okonomiyaki* (a type of Japanese pancakes). We kept the British expression in the previous example, but did not keep the German or Japanese in the data sets. In the paid crowdsourcing process we believe that the formality of the task as well as the nature of the task might have led people to opt for a more standard version of English. In the game-based approach people were

[11] www.yago-knowledge.org/

neither informed nor aware of the fact that all other participants were equally based in Edinburgh. Since participants met online without any means of identifying the other player in each game, virtually no local expressions could be found in the final dataset. However, participants presumed to be in the same region since they repeatedly used cardinal directions to describe cities, such as *West*, which only helps in the game if this direction points to the same locations for both players.

5 Discussion

The two crowdsourcing techniques that we used in this approach returned useful input for the ontology building process. We found that the time needed to obtain data from the mechanized labour-based approach exceeded the time of the game-based approach to return the same amount of data. 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. This is because the incentive to participate in an interactive online game seemed much higher. In fact, participants asked for the permission to play again after the first session, and four of the 30 participants joined a second session. One crucial point in accelerating the labour-based approach was the overall contributor satisfaction. In the pilot study, which also served to evaluate the accuracy of our test questions, we obtained a lower ranking of only 3.5 out of 5 points and several comments from contributors regarding suggestions to improve the quality of the test questions. In the actual study, we obtained a rating of 4.2 out of 5 points, which considerably increased the frequency with which new participants joined the task. Thus, we could observe that a high ranking can considerably accelerate the labour-based approach, however, it is unlikely that it even a the highest possible ranking would come close in performance to the game-based approach.

The low number of and locally restricted recruitment method of participants for the game-based approach might have a biasing effect on the data set obtained from this method. While the results from the concept formation stage seem to be comparable to the data obtained from the labour-based crowdsourcing step, we still believe that it would be more interesting to compare both methods with a larger number of participants. To this end, we have already developed a mobile app version of GUESSence that will allow an asynchronous access to the game and allow for an open access without any restriction to a specific region.

When applying the ontology building process to the data sets obtained with both crowdsourcing techniques, it can be observed that the mechanized labour-based technique returns more specific categorizations. This implies that the game is not useful as an extension of the first method, as we initially suggested. However, there are only minor differences in the results obtained with both approaches, and they both perform well when compared to our gold standard ontology. This means that the game returned results that were as robust as the direct technique, and can therefore be used on its own to retrieve descriptions. These kind of describing and guessing games are very popular, easy to play, and can be extended to different domains, which turns them into a good candidate to obtain this kind of data. They can even be included in an online gaming platform, which would provide very large sets of data.

This extensibility of the game-based acquisition method to other domains or other target concepts than *city* is also true for the labour-based approach. Any instances of a specific concept could be used to trigger responses from participants as long as they are general enough to be known to a crowd of non-experts, such as *Food* as a general category or *Movies*. The results of the first approach could again be used as taboo words for the game. Alternatively, the taboo words could be generated manually. However, the design of the game relies on those input restrictive characterizations we call taboo words. The main restriction in terms of portability of the approach to other domains are: a) the domain and its instances have to be known by the participants to a degree that allows them to describe them, b) the domain has to be covered by some predefined knowledge or lexical resource to allow for the disambiguation and classification approach as it is proposed here to work. This means that highly domain-specific contents, such as *fishing rod* and the specific types of rods as instances, would not work well for this approach since a low number of participants would be able to describe their characteristics and few resources would contain the words, however, the more general category *sport* or *sport equipment* could be used.

Regarding the concept formation, the two resources that we used for the extraction of senses (Wordnet and Word Reference) were accurate in that they contain 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 semantic category, 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 labelling 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 comparison of the categories retrieved from Word Reference with the ones in the gold standard shows that in most cases the labels match. Some of the ones that do not match directly are in turn 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 approach, since players tend to describe instances not only with their properties but also with properties from their upper categories.

6 Conclusions

We presented and discussed methods to automatically build concept descriptions from crowdsourced data. The two crowdsourcing techniques that were proposed gave good results in terms of quantity and reflecting associative knowledge. The results obtained with the game-based approach are as robust as the ones obtained with the mechanised one, although slightly less fine-grained. The final results we obtained could be very valuable as a seed for ontology learning to be extended with hierarchical, non-hierarchical relations, and axioms.

Multiple directions of research are derived naturally from

this work. Regarding the crowdsourcing methods presented, it would be interesting to compare, for this particular problem, the use of implicit techniques, like the ones we propose, with explicit ones. The implicit techniques already have as an advantage that they can be easily presented as a game, making the task more attractive. However, it would be interesting to compare the differences in the results. To this end, a third experiment in which users are asked directly to name properties of cities should be performed.

At the moment we focus on concept formation. In order to have more interesting and useful ontologies, this part should be developed to extract hierarchical and more informative non-hierarchical relations. There is a variety of approaches that tackle the relation extraction problem, both with automated and crowdsourcing techniques. However, their adequacy to our problem should be analysed, since they are not particularly designed to identify relations between a concept and its attributes.

ACKNOWLEDGEMENTS

This research has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 607062 /ESSENCE: Evolution of Shared Semantics in Computational Environments/.

References

[1] Luigi Atzori, Antonio Iera, and Giacomo Morabito, 'The internet of things: A survey', *Computer networks*, **54**(15), 2787–2805, (2010).

[2] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro, 'An enhanced lesk word sense disambiguation algorithm through a distributional semantic model', in *COLING*, pp. 1591–1600, (2014).

[3] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Fedelucio Narducci, 'Uniba: Exploiting a distributional semantic model for disambiguating and linking entities in tweets', in *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015), CEUR Workshop Proceedings 1395, CEUR-WS.org*, eds., Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie, p. 62, (2015).

[4] Brandon Bennett, 'What is a forest? on the vagueness of certain geographic concepts', *Topoi*, **20**(2), 189–201, (2001).

[5] 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 Resources*, pp. 101–108, (2004).

[6] Steven Bird, 'NLTK: the natural language toolkit', in *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72, (2006).

[7] 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 Sciences (HICCSS)*, pp. 2289–2297, (2012).

[8] Philipp Cimiano, 'Ontology learning from text', in *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, ed., Philipp Cimiano, 19–34, Springer, (2006).

[9] Philipp Cimiano, Andreas Hotho, and Steffen Staab, 'Learning concept hierarchies from text corpora using formal concept analysis', *Journal of Artificial Intelligence Research (JAIR)*, **24**(1), 305–339, (2005).

[10] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy, 'Crowdsourcing systems on the world-wide web', *Communications of the ACM*, **54**(4), 86–96, (2011).

[11] 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*, pp. 139–148, (2010).

[12] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, 'Incorporating non-local information into information extraction systems by gibbs sampling', in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370, (2005).

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

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

[15] So Young Kwon and Lauren Cifuentes, 'The comparative effect of individually-constructed vs. collaboratively-constructed computer-based concept maps', *Computers & Education*, **52**(2), 365 – 375, (2009).

[16] Egoitz Laparra, German Rigau, Piek Vossen, et al., 'Mapping wordnet to the kyoto ontology.', in *LREC*, pp. 2584–2589, (2012).

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

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

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *ICLR 2013*, (2013).

[20] Peyman Nasirifard, Slawomir Grzonkowski, and Vassilios Peristeras, 'Ontopair: Towards a collaborative game for building owl-based ontologies', in *Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb), Workshop at the 5th European Semantic Web Conference (ESWC08)*, (2008).

[21] 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, (2013).

[22] Iuliana-Elena Parasca, Andreas Lukas Rauter, Jack Roper, Aleksandar Rusinov, and Guillaume Bouchard Sebastian Riedel Pontus Stenetorp, 'Defining words with words: Beyond the distributional hypothesis', *ACL 2016*, 122, (2016).

[23] 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).

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

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

[26] 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, Berlin, Heidelberg, (2007). Springer.

[27] 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).

[28] 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, (2013).

[29] Luis Von Ahn, Mihir Kedia, and Manuel Blum, 'Verbosity: a game for collecting common-sense facts', in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 75–78. ACM, (2006).

[30] 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).

[31] 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).

[32] 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).