

# Expert-sourcing Domain-specific Knowledge: The Case of Synonym Validation

Michael Unterkalmsteiner  
Software Engineering Research Lab Sweden  
Blekinge Institute of Technology  
Karlskrona, Sweden  
michael.unterkalmsteiner@bth.se

Andrew Yates  
Max Planck Institute for Informatics  
Saarbrücken, Germany  
ayates@mpi-inf.mpg.de

## Abstract

One prerequisite for supervised machine learning is high quality labelled data. Acquiring such data is, particularly if expert knowledge is required, costly or even impossible if the task needs to be performed by a single expert. In this paper, we illustrate tool support that we adopted and extended to source domain-specific knowledge from experts. We provide insight in design decisions that aim at motivating experts to dedicate their time at performing the labelling task. We are currently using the approach to identify true synonyms from a list of candidate synonyms. The identification of synonyms is important in scenarios where stakeholders from different companies and background need to collaborate, for example when defining and negotiating requirements. We foresee that the approach of expert-sourcing is applicable to any data labelling task in software engineering. The discussed design decisions and implementation are an initial draft that can be extended, refined and validated with further application.

## 1 Introduction

The training and validation of natural language processing models, that are based on supervised machine learning, require data that is labelled by humans. Creating labelled data, in particular if it is domain specific, is costly and can require expert knowledge. Furthermore, the lack of high-quality labelled data may prevent the transfer of an approach from one domain to the other, simply because not enough labelled data exists to train the model [Fer18].

Crowdsourcing platforms provide the possibility to harvest human intelligence that can be used for data labelling. While this works well for tasks that target the humans' predisposition for pattern recognition, tasks for which domain-specific knowledge is required cannot be outsourced to an arbitrary crowd. Such tasks need to be designed such that a limited target group remains engaged with the data labelling task and experiences benefits from participation. In this paper, we provide some insight in an ongoing study and provide motivation for the design decisions we made when adopting an existing crowdsourcing tool for our particular task: validation of domain-specific synonym candidates.

## 2 Background

Our current research focuses at supporting requirements engineers to adopt an object classification system, CoClass<sup>1</sup>, from the construction business domain. The classification is planned to be used throughout the organization to identify and trace specified, designed, constructed and eventually maintained objects. CoClass is a hierarchical ontology of construction objects that provides a coding system, a definition and synonyms for each object. CoClass is still under development and many object to synonym mappings are still incomplete. These mappings are however important for the use of the classification system as it allows users, with different background and vocabulary, to find the objects they are looking for. Furthermore, we plan to use the ontology to automatically classify natural language requirements such that they can be traced during the life-cycle of a project.

### 2.1 Domain-specific synonym detection

In order to fill the synonym gaps in CoClass, we use a learning-to-rank approach for domain-specific synonym detection [YU19]. The basic idea of this supervised approach is to learn term associations from a domain specific corpus, using features that indicate the synonymous use of a term. The approach produces a list of synonym candidates for each term defined in CoClass (1430 terms with each 1000 synonym candidates). A preliminary evaluation of the candidates with a domain expert suggests that only  $\sim 1\%$  of the synonym candidates are true synonyms (10 in 1000). While this precision might seem underwhelming, automated synonym detection *is* difficult and should be compared against its manual alternatives or evaluated against the cost of not discovering new synonyms at all.

## 3 Expert-sourcing synonym validation

Reviewing 1,430,000 synonym candidates would be a monumental task for an individual. While crowdsourcing [DRH11] the task to the general public would be possible, it would likely not succeed, as the task is language (Swedish) and domain (construction business) specific, limiting the potential and reliable participants considerably. We chose therefore to use a crowdsourcing framework, Pybossa<sup>2</sup>, that allows us to control all aspects of the validation process: participants, data storage and task design. Pybossa provides important infrastructure for realizing a crowdsourcing project, such as task importing, management, scheduling, and redundancy, user management and results analysis. In addition, Pybossa provides a REST API and convenience functions that can be used for tasks, e.g. a media player for video/sound annotation tasks or a PDF reader for transcription tasks. In the remainder of the paper, we focus on the task design and the decisions that were made in order to make the validation process efficient and effective. The code for task presentation and analysis is available online<sup>3</sup>.

### 3.1 Task design

The validation task is separated into two phases. In phase 1, the selection, the expert selects 0.. $n$  synonyms from a list of candidates for a particular target term. In phase 2, the result, the expert receives feedback on his/her selection. Screenshots of the respective phases are shown in Figure 1 and 2. The red markers are inserted for referencing purposes, used in the following discussion.

Panel 1 in Figure 1 shows the target term for which the expert needs to select synonyms. In this area, we show also the hierarchical structure of CoClass under which the target term (transl.: fence) can be found (transl.: Components  $\gg$  Limiting objects  $\gg$  Access-limiting objects  $\gg$  Fence), including the coding that is used for such objects (R  $\gg$  RU  $\gg$  RUA). We also show the definition used in CoClass of the target term (transl.: access restricting object by a horizontal elongated barrier with a vertical extent). The purpose is to provide context, to foster organizational learning [Kim93] and to develop a common vocabulary that potentially reduces misunderstandings in the organization.

Panel 2 in Figure 1 shows the list of candidate synonyms. We group candidate synonyms with affinity propagation clustering [FD07], measuring similarity with the Levenshtein distance. This reduces the perceived number of terms an expert has to inspect as similar terms can be accepted/rejected in one task. If the expert is

---

<sup>1</sup><https://coclass.byggjant.se/en/about#about-coclass>

<sup>2</sup><https://github.com/Scifabric/pybossa>

<sup>3</sup><https://github.com/munterkalmsteiner/pybossa-trafikverket-theme>

1

## Select the correct synonym(s)

Komponenter >> R:Begränsande objekt >> RU:Tillträdesbegränsande objekt >> RUA:Barriär

**Target term: Barriär**

Definition: tillträdesbegränsande objekt genom en horisontell långsträckt barriär med vertikal utsträckning

---

2

### Candidates:

- ytvattenförekomst
- råvattentäkt
- naturvårdverk
- livsmedelsverk
- stängsel
- parkeringsplanka
- ytvattenverk

Only press skip if you are not sure. If none of the candidates is a synonym, do not check any term and submit.

---

3

You are working now on task: 1289

You have completed: 7 tasks from 26

Figure 1: The selection phase of the task

Definition: tillträdesbegränsande objekt genom en horisontell långsträckt barriär med vertikal utsträckning

4

### Results:

Candidate	Synonym <sup>⓪</sup>	Alignment <sup>⓪</sup>
ytvattenförekomst	—	4 0
råvattentäkt	—	4 0
naturvårdverk	—	4 0
livsmedelsverk	—	4 0
stängsel	✓	2 2
parkeringsplanka	✗	2 2
ytvattenverk	—	4 0

**1** Synonym

Correctly identified synonym: ✓

Missed actual synonym: ✗

New synonym: ★

No synonym: —

**2** Alignment

This column indicates how well you agree with your colleagues. The numbers show how often you **agree** or **disagree** in total, on the assessment of the particular synonym candidate.

You are working now on task: 1289

Figure 2: The results phase of the task

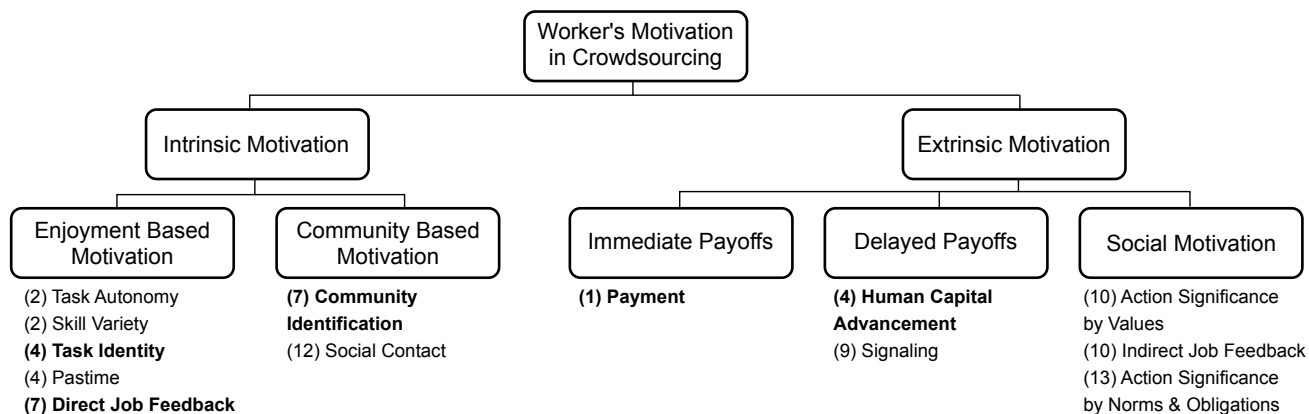


Figure 3: Model for worker’s motivation in crowdsourcing, adapted from Kaufmann et al. [KSV11]

not sure about the meaning of the term or the synonym candidates, (s)he can skip the task and proceed to the next one. Panel 3 in Figure 1 shows the overall progress, i.e. tasks done of the total number of tasks.

Once the expert has made a decision, the results for the particular task are stored and analysed in order to provide immediate feedback to the expert. An example of the analysis is shown in Panel 4 in Figure 2. In the second column of the results table, we show whether the selected term is a correctly identified or a missed actual synonym, according to the already defined synonyms in CoClass, or a completely new identified synonym. In the third column, we show how well aligned the current expert is with other experts that have already performed the same task. For example, the selection of the expert in Figure 2 has missed the actual synonym “parkeringsplanka”, and so did another user. They agree that “parkeringsplanka” is not a synonym of “barriär”. However, two other experts had a different opinion, i.e. “parkeringsplanka” is indeed a synonym of “barriär”. Once the tasks are completed, it is straightforward to identify new synonyms with a simple majority vote.

### 3.2 Motivational aspects of expert-sourcing

When we designed the task, we considered how to create a win-win situation for the participating experts, management which pays for the time spent on the task, and researchers. There exists some evidence that intrinsic motivation is more important than extrinsic motivation for crowd-sourcing workers [KSV11]. Figure 3 shows different aspects of motivation and their relative importance ranking (1-13), based on a survey of 431 Amazon Mechanical Turk workers. In the remainder of this section, we discuss our strategies to foster some aspects of worker motivation.

The synonym selection task should transfer some knowledge to the participants. We provide that by showing term definitions, the CoClass hierarchy and code under which the term is found. This fosters individual learning as well as organizational learning as it promotes a common vocabulary (Human Capital Advancement, i.e. motivation to enable training of skills). Similarly, the feedback on the results page helps individuals to understand how well they are aligned with their colleagues (Direct Job Feedback, i.e. motivation provided by the perception of achievement; Community Identification, i.e. the subconscious adoption of norms and values). For management, this could also be useful information as it could indicate where adjustments in documentation or training are needed. Since we know exactly how much time each expert has spent on their tasks, we can quantify the cost for collecting synonyms and potential terminology misalignments (Payment, i.e. motivation by monetary compensation). Such figures can help to get management buy-in when extending the study or replicating it in another organization.

A potential threat to the validation of the synonyms is the result page where we show the alignment of experts immediately after their choice (Task Identity, i.e. the extent to which a participant perceives that his/her work leads to a result). Therefore, we randomize the presentation of tasks (in blocks of five, i.e. after five tasks we change the target term), counteracting conscious or unconscious bias. Finally, we seed a true synonym if the expert did not select a synonym after 10 tasks in a row. The intention is both to keep the participant motivated by “finding” a synonym and to verify that the expert is still paying attention to the task and not submitting random answers.

In Figure 3, we highlight in **bold** typeface which motivational aspects we address. We briefly discuss which aspects are not covered. Task Autonomy refers to the degree to which creativity and own decisions are permitted

by the task. The nature of data labelling tasks leaves little leeway and creativity would rather be counter productive. It would be difficult to design a task that caters for this motivational aspect. Skill Variety refers to the usage of different skills for solving a task that match to the available skill set of the worker. One way to address this motivational aspect would be to segment the CoClass terms into themes that require specialized subdomain knowledge, matching a subset of participants' specialized background and expertise. Pastime refers to the motivation to do something in order to avoid boredom. One could argue that, since the synonym selection task can be performed on mobile devices (e.g. while riding the train to work), this motivational aspect is covered. On the other hand, the task *is* work and part of the professional activities of an employee, making this motivational aspect not applicable to our context. We do not address any aspects from the range of social motivations. Indirect Job Feedback, i.e. motivation through feedback about the delivered work, for example through comments and other encouragements could however be implemented. Finally, we do not use yet any form of gamification mechanisms. Leaderboards and level systems can be effective means to increase long-term engagement and quality of output [MHK16].

## 4 Conclusions and Future Work

In this paper, we suggest expert-sourcing as a mean to acquire labelled data from domain experts. We illustrate the adoption of a crowd-sourcing platform and the design of the data labelling task for domain-specific synonym identification such that it is engaging and useful for the participating experts. We are currently in the process of piloting the approach with select domain experts and gather feedback on the task design. Once the task design is stabilized, we intend to deploy the data collection mechanism to approximately 500 participants.

While we apply the approach to a narrow, specialised, problem (synonym identification), the idea and design decisions to cater for motivational aspects are generally applicable to any data labelling task in Software Engineering. One could design tasks to evaluate the quality of certain artefacts and use this assessment to train a classification algorithm, for example to evaluate the degree of ambiguity in statements of requirements specifications, the understandability of test cases, identification of code refactorings, detection of code smells or the readability of source code.

## References

- [DRH11] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Commun. ACM*, 54(4):86–96, April 2011.
- [FD07] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, February 2007.
- [Fer18] Alessio Ferrari. Natural Language Requirements Processing: From Research to Practice. In *Proceedings 40th International Conference on Software Engineering*, pages 536–537, Gothenburg, Sweden, 2018. ACM.
- [Kim93] Daniel H. Kim. The Link Between Individual and Organizational Learning. *Sloan Management Review*, 35(1):37+, 1993.
- [KSV11] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems*, volume 11, pages 1–11, Detroit, Michigan, USA, 2011.
- [MHK16] Benedikt Morschheuser, Juho Hamari, and Jonna Koivisto. Gamification in crowdsourcing: a review. In *Proceedings 49th Hawaii International Conference on System Sciences (HICSS)*, pages 4375–4384. IEEE, 2016.
- [YU19] Andrew Yates and Michael Unterkalmsteiner. Replicating Relevance-Ranked Synonym Discovery in a New Language and Domain. In *Proceedings 41st European Conference on Information Retrieval*, Cologne, Germany, 2019. Springer.