# Effectiveness of Automatic and Semi-Automatic Methods to Collect Common Sense Knowledge

**François Ezard**
**Supervisor(s): Gaole He, Ujwal Gadiraju, Jie Yang**
**EEMCS, Delft University of Technology, The Netherlands**
22-6-2022

**A Dissertation Submitted to EEMCS faculty Delft University of Technology,**
**In Partial Fulfilment of the Requirements**
**For the Bachelor of Computer Science and Engineering**

## Abstract

Common sense knowledge (CSK) comes naturally to humans, but is very hard for computers to comprehend. However it is critical for machines to behave intelligently, and as such collecting CSK has become a prevalent field of research. Whilst a lot of research has been done to develop CSK acquisition methods, not much work has been done to survey the literature that already exists. Furthermore the surveys that have been done are outdated, and as such there is a clear gap in the literature. This paper will survey the different approaches to CSK acquisition and evaluate their effectiveness, as a way of gauging their real life applicability. It will also compare the current *state of the art* methods, to some previous work to illustrate the progress that has been made and project that into the future. Furthermore this paper will also create a taxonomy categorizing the surveyed literature in order give a better overview of existing methods. Finally, from the literature surveyed it is clear that these methods have made a lot of progress, but aren't quite yet at the same level as human performance. Nevertheless they have become robust enough to be deployed in real applications.

## 1 Introduction

Common sense knowledge (CSK) can be defined as the set of knowledge that all ordinary humans are assumed to have and comes naturally to humans [3]. CSK encompasses a wide variety of knowledge, that is acquired through every day experiences. This can be something as simple as "Fire burns" or "lemons are sour". This type of knowledge is what gives humans the ability to behave intelligently [3]. It is for that reason that researchers have been looking to incorporate CSK in machines. Unfortunately although CSK comes naturally to humans, it is very difficult for computers to understand it and even more so reason about it [31].

In the realm of computer science a program is said to have common sense if "it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" [28]. Although not evident at first, CSK is critical in the field of Artificial Intelligence (AI) [29, 34], as it makes machines more intelligent, bringing them closer to human-level intelligence [8, 47]. One of the many applications of CSK in AI is for question answering, as it enables agents to make educated decisions when choosing between answers [34]. In fact, in the field of Natural Language Processing (NLP) it is widely accepted that computers can't fully understand text without CSK [32]. CSK helps humans understanding text better than machines, as they can infer knowledge that isn't explicitly written in text which could help understand the context [30]. Incorporating CSK in machines would also make it so that machines behave in a way that is closer to the expectations of humans [43]. For those reasons CSK is one of the pillars of computers reaching human levels of intelligence [15].

The issue is that CSK tends to be implicit [19], and thus is rarely articulated in communication [3, 4, 7], making it difficult for computers to recognize it and extract it. This can cause automatic methods to miss basic relationships which leads to unreliable results [20]. As a result even though the importance of CSK has been clear for a while, CSK acquisition has been the bottleneck in the development of intelligent systems [19, 43]. For that reason there has been a lot of research done in this field over the past few years [25, 35].

Several approaches have been used in an attempt to solve the problem of CSK acquisition. The most obvious approach is to have humans manually input CSK in a machine-friendly form, since CSK is simple for humans. Although such approaches have good precision, they lack coverage due to the sheer amount CSK that exists [23, 36, 47]. As such automatic and semi-automatic methods are more desirable due to their scalability.

This paper surveys the literature relating to automatic and semi-automatic methods, and evaluates their effectiveness. Furthermore this paper will categorize the different methods in order to give a better overview of the different approaches to solving the problem of CSK acquisition. The aim of this paper is to evaluate the performance of methods that are currently *state of the art*, and evaluate how applicable they are to real world problems. Although similar surveys have been done in the past, they are outdated in a field that moves rapidly.

There are many challenges in making such a survey paper, the first one being the amount of papers published. The relevant papers then have to be identified from the cluster of collected literature. After having read over the majority of the relevant papers, an initial taxonomy can be created ordering methods based on the similarity of their approach. Creating the taxonomy is a continual process as it constantly needs to be tuned to best fit the literature that is being gathered. Comparing the literature is also a big challenge as we have to rely on the individual evaluations made (if there was one) and there is no guarantee that these evaluations are comparable.

The section right after this introduction will discuss the methodology of this project. It will give an insight into the steps taken, as a way to ensure this research is reproducible. Section 3 will introduce the taxonomy, the process of creating it and explain the characteristics of the different sub-categories. Section 4 will focus on answering the core of the research question, as it focuses on the effectiveness of the methods and the metrics used to evaluate it. Section 5 will discuss the results and their significance. After that 6 will re-iterate the main talking points, before concluding the findings of this paper. Section 7 is the last section and is devoted to the ethics involved in this paper.

## 2 Methodology

This section will discuss the process that led to collecting the papers referred to in this paper, as well as the process of writing this paper. It will follow the logical order of the project with each subsection delving into further detail of a specific aspect.

### 2.1 Collecting Literature

The supervisors supplied us with anchor papers from which to start our research from. For this project the anchor papers were "COMET : Commonsense Transformers for Automatic Knowledge Graph Construction" [10] and "Commonsense Knowledge Mining from Pretrained Models" [17]. Naturally these papers were the starting point for the research, as they helped familiarize with the topic and pointed to further relevant papers. More specifically by looking at the work that was cited in those papers as well the related work section. Having a quick read of the abstract and introduction gave an indication of which of those papers would actually be useful for this project. That way some of the papers were discarded without wasting time, whilst the others were stored in Mendeley Reference Manager [1].

Of course solely relying upon this to gather literature is not sufficient, as it would not only limit the scope of the research but also papers would get more and more outdated, as a paper can only cite an older paper. For that reason we also used Google Scholar[2] to search for other papers. This search was not only to find papers proposing more methods, but also to find survey papers on the topic. Finding new methods was done by using keywords such as "automatic" in combination with "common sense knowledge", if we're looking for automatic methods. To find surveys on the other hand we use keywords such as "Survey" or "Overview" in combination with "common sense knowledge". Finding survey papers serves multiples purposes, firstly it gives an indication of what a survey paper has to offer and the structure of such a paper. Secondly, it points to further methods that are used, which can then also be used in this survey and point to further literature themselves. Lastly, it gives an indication of how to compare different methods against each other and evaluate individual methods. Looking at the evaluation section of a paper gives an indication of how the authors evaluate their own work, but a survey shows how others evaluate their findings and compare it with other papers.

### 2.2 Analyzing Literature

As stated in the previous subsection, Mendeley Reference Manager [1] was used to store all of the literature gathered. The reason for choosing this platform is that it provided several tools that were critical for storing and reading literature. Firstly when reading a paper, we highlighted all the relevant parts, so that it saves time when reading over it later on. Secondly if simply highlighting wasn't clear enough, we also added comments to clarify what is relevant about that specific section of text. Comments are more flexible than simply highlighting, as they can give further detail beyond just indicating that a section of text is important. Lastly, and most importantly, Mendeley allows to assign tags to a document. These tags are on the entire document and can be seen without having to open the document. Furthermore it is possible to filter papers in a collection based on their tags. This was particularly useful, as it allows to assign tags to papers depending on what they are useful for, and save time when looking for a paper relating to a specific topic or section. We started with tags that were broader, such as "Introduction", "Automatic" or "Evaluation", and gradually made them more precise. As the project progressed and the taxonomy became more specific, it became clearer what sections would be needed and as such tags such as "Reasoning Acquisition" or "Motivation" were introduced instead of "Automatic" and "Introduction" respectively. It is worth noting that the tags weren't created in advance and simply assigned, but rather that they were created to best describe a paper and aggregated later. These tags were instrumental in creating the taxonomy but more information regarding that process in section 3. Tags were also useful to give an indication of what was still missing and thus helped prioritize work.

[1] www.mendeley.com/reference-manager

[2] scholar.google.com

## 3 Taxonomy

This section is the start of the core of this paper, as it introduces the taxonomy and the methods considered. The taxonomy is based on the approach used to collect the knowledge. Note that not all the methods categorized in the taxonomy will be discussed, so refer to appendix A in order to see an overview of all the methods and the paper they came from. As can be seen in appendix A some subsections only have one method in it, which might seem odd. The reason for those sub-categories still existing, is that we felt that there was a clear theoretical gap in which more methods could fit into. This means that even though at the moment there aren't many methods using that approach, we felt as though there could be more in the future.

### 3.1 Overview

Creating a taxonomy is critical for this paper, as automatic & semi-automatic methods are broad categories. Creating these sub-categories highlights which methods are closely related and gives room for further analysis. Methods can then be compared within their category and categories can be compared to each other. The taxonomy developed can be seen in Figure 1.

In this taxonomy the automatic branch has more depth than the semi-automatic one due to the fact that there is more literature that falls under that branch. As such there are more types of methods and there is a greater need to create sub-categories to distinguish them.

Figure 1 shows the final taxonomy created. The categories in the taxonomy are not arbitrary. The process of deriving the sub-categories of the taxonomy was a continuous process that used an open coding approach. The first step was to gather some literature and reading it over to understand the approaches conceptually. From that we created a few tags per method, to describe the approach of that method. After having done that for all method we had a look at tags that came up several times to create a first draft of the categories. The way in which the categories describe the methods it encompasses can vary, it can be due to the goal of the methods, such as Knowledge Base (KB) Completion methods which aim to improve coverage. It can also be based on the input the methods take, such as the difference between

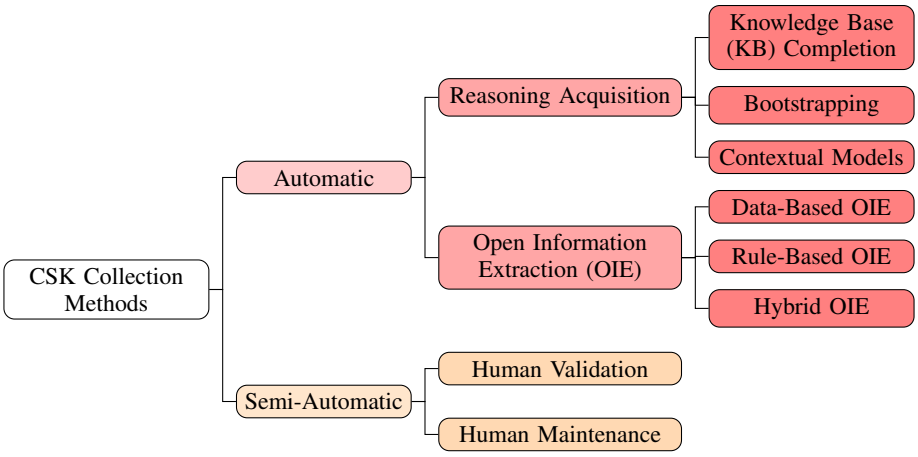


Figure 1: Taxonomy of Knowledge Acquisition Methods

Open Information Extraction (OIE) methods and Reasoning Acquisition methods. The way in which the categories are divided took some judgement and was a long process. As more literature was gathered we had to see under which existing category it would best fit or change the existing categories to make it fit. Certain methods could've been categorized in different ways, as there was some overlap, and it took some judgement to decide under which category a method best fits. Sometimes this was also an indication that the architecture of the taxonomy had to be re-thought. It is worth noting that not all the methods that exist can be categorized in this taxonomy, as some are outside the scope of this project. To see all the methods and their sub-category refer to appendix A.

### 3.2 Related Work

Before discussing the taxonomy and the methods it is important to introduce some key papers published in the field of CSK acquisition. Although these papers are outside the scope of this project, they help understand the literature that is relevant and provide context. Some of these papers notably introduced some of the biggest KBs and as such are used by a lot of other methods as a way to either train or test their model.

The Cyc [22] project is one the oldest and largest CSK bases. Cyc was started in 1984 with the goal of codifying CSK in a machine-friendly way, in order to enable human-like reasoning from computers [47]. Originally the knowledge was only entered by trained engineers using a specific language called CycL [20]. However there have been other methods used since then, such as using untrained volunteers or extracting CSK from the web.

The Open Mind Common Sense (OMCS) project [39] is one of the largest CSK bases out there. OMCS supports several languages, but English is the language that has the most statements. The knowledge is collected manually, by having untrained volunteers do the bulk of the work. This is a way of making the process slightly more scalable, as it means more people are able to do the job.

ConceptNet [20] is a network representation of the CSK collected in majority by the OMCS project. In ConceptNet the nodes represent concepts whilst the edges represent relations between concepts, and as such two nodes with one edge represents an assertion [47]. There have been multiple revisions since the original release of ConceptNet back in 2002.

**AT**las **O**f **M**ach**I**ne **C**ommonsene or ATOMIC [37], is a crowdsourced knowledge graph that focuses on "if-then" relations [41]. Experimental results has shown that incorporating the structure of "if-then" relations leads to more accurate inference [37].

Another category worth mentioning are games with a purpose (GWAP). The idea behind those is to have humans take part in games where they perform tasks that are simple for humans but difficult for machines [11, 12]. Embedding knowledge acquisition in a game is supposed to create a win-win situation, as it entertains the gamers and provides useful CSK [11, 12]. GWAPs are a very powerful category of methods, but are outside the scope of this project, refer to my teammate Ilinca Rentea's paper instead.

### 3.3 Automatic: Reasoning Acquisition

Reasoning Acquisition refers to methods that automatically infer CSK from a pre-existing KB [47]. This means that there already has to be an existing KB for such methods to work, and that they can't just be given text as input. This might seem odd at first, as how could this approach create any new knowledge, but this is actually a very powerful approach. Because CSK is rarely stated in order to grow the domain of a KB there has to be another source of CSK other than text [4].

**Common Sense Knowledge Base Completion (CSKBC)**
Common sense knowledge base completion (CSKBC) methods have the goal to improve the coverage of a KB [23]. Usually the nodes of the graph are kept the same, and the method only tries to predict edges between these nodes [21]. Such methods usually perform well in terms of precision, but producing new knowledge is more difficult. There are several approaches to achieve this goal, but a common issue

is that statements end up being slight re-wordings of the ones present in the training set [21].

**COM**ens**E**nse **T**ransformers, is a method that was introduced in "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction" [10], a paper published in 2019. COMET trains a generative model transformer, in order to infer new relations from a knowledge graph [41]. It uses an existing KB to learn to generate phrases and it has the ability to create new nodes in the knowledge graph [10]. COMET was trained and tested on two different domains, one was the ATOMIC [37] KB and the other was the ConceptNet3 [20] one.

Memory Comparison Network (MCN) is a method that was introduced in a paper titled "Leveraging knowledge bases for future prediction with memory comparison networks" [2], published in 2018. The idea behind this method is to use the temporal relations from a KB in order to predict future unseen events, to increase coverage. MCN has a very special functionality, as it provides an explanation for its predictions. This is helpful to convince humans of the reasoning and also helps indicate when the rules in the KB are insufficient [2].

"Commonsense Knowledge Base Completion" [23] is a paper that was published in 2016. This paper introduces several different methods, that fall under two main categories: Bilinear Models and Deep Neural Network (DNN) Models. The downside to Bilinear Models is that the size of relation matrices grows quadratically and as such it slows training and requires more data to train. Also this creates restrictions on how terms can interact, which DNNs don't do.

**Contextual Models**

Methods that fall under this category don't rely upon hand-crafted features or KBs [31], although they do have some overlap with CSKBC methods. These methods make use of pre-trained language models, and tune them to make them work for CSK acquisition. The difference is that CSKBC methods specifically aim to increase coverage, whilst contextual models rely on a pre-trained language model to infer CSK, but not purely to increase coverage.

**A**dversaria**L** training algorithm for commonsense **I**nferen**CE** (ALICE) is a method that was introduced in "Adversarial Training for Commonsense Inference"[31] a paper published in 2020. It makes use of the RoBERTa model, and relies upon the true labels and model predictions to infer CSK. It also makes use of adversarial training, which is the idea of disturbing the distribution of the data in the embedding space, as a way to prevent over fitting and generalize better.

"Stagewise Fine-tuning BERT for Commonsense Inference in Everyday Narrations" [24] is a paper published in 2019 that introduces a method that falls under this category. It is important to note that the method introduced was created for a machine comprehension task, that tested a system's ability to answer questions about text [30]. As indicated by the name of the paper, this method uses BERT [14] as their pre-trained language model. There are two stages in which this model is tuned, in the first stage it is tuned on additional datasets to learn more CSK and in the second stage it is tuned for the target task.

**Bootstrapping**

Methods that take a bootstrapping approach to CSK collection, often start with sparse amounts of data and continuously use their data to learn. Bootstrapping methods work in a fully automatic manner, but rely on having some seed CSK. The common characteristic of these methods is the bootstrapping phase, which refers to the beginning where the seed CSK is used to create new CSK. That new CSK is then used as part of the seed to create new CSK and so on. It is important to note that predicates should not be predefined, as this limits the potential of the model since it means they don't have the ability to identify unknown predicate structures [45].

ASTRID is a method that was introduced in a paper titled "ASTRID: Bootstrapping Commonsense Knowledge" [45] published in 2021. The inspiration for this method comes from the way children learn about the world. Before children are able to communicate they already have experiences which creates knowledge, but that knowledge is not labeled until later when they have they are able to communicate. From there children are innately curious which leads them to continuously ask questions until they have acquired enough knowledge to make their own conceptual connections.

### 3.4 Automatic: Open Information Extraction (OIE)

Open Information Extraction (OIE) methods are fully unsupervised methods that don't have any knowledge of the type of entities they'll be mining from [1, 6]. These methods are able to do so from text, which can come from a variety of sources but a lot of methods like to use the Web as a source as there is an enormous amount of text available. The issue however with methods that use the web as a source is that they tend to suffer from confusion and inconsistencies [15].

In order to extract CSK from text, OIE methods make use of several natural language processing (NLP) techniques [47]. As stated previously, CSK is rarely explicitly stated in natural language [17], but it is still present some of the time or algorithms focusing on lexical patterns to extract CSK wouldn't exist [46]. All OIE make use of a set of patterns, but the difference lies in how these patterns are identified [1]. The main strength of OIE methods are that they are extremely scalable as they only need to read over the text a constant amount of times and can do so quickly [5, 16]. Another benefit of such methods is that they don't depend on any pre-existing KBs and as such can be readily deployed on a new domain.

**Data-Based OIE**

As previously stated, the difference between OIE methods lies in how the extraction patterns arise. For data-based OIE methods those patterns are automatically generated from the training data [6]. The advantage of such approaches is that they are easily to deploy as no heuristics have to be manually created, which can be time consuming. However the training data has to be representative and even if it is the patterns often don't work as well as handcrafted ones.

"Open Information Extraction from the Web" [5] is a paper

that was written in 2007, and introduced the OIE paradigm. This paper also introduced TextRunner, representing the first generation of OIE methods for CSK acquisition and falls under the sub-category of data-based OIE. TextRunner works by making a single pass over the input text, in which it gives tags to a each word with their most likely role in the sentence. This is used to identify noun phrases and as such eliminate non-essential phrases. Each of the remaining noun phrases has a probability of belonging to an entity used to construct candidate tuples that are then presented to the classifier. Depending on whether or not the classifier views the tuple as credible, it will either discard it or store it. The performance of TextRunner is unimpressive by today's standards, but showed a lot of promise at the time. For that reason and the fact that it was the first generation of OIE methods, it is often referred back to in more recent work.

"Commonsense Knowledge Mining from the Web" [46] is a paper published in 2010, which introduces a prototypical data-based OIE method. The algorithm introduced uses a seed set of CSK to build their classifier from. Then they use that classifier when reading the text in order to evaluate the quality of the new CSK. Whilst reading the text the algorithm also continues to induce new patterns meaning it is continuously learning. Similarly the classifier also uses the new CSK in order to continue to learn [46]. This method highlights the ability of data-based OIE methods, to continuously learn and adapt to the domain they are deployed on.

**Rule-Based OIE**

Rule-Based OIE methods rely upon manually crafted rules or heuristics in order to extract CSK [6, 8]. Handcrafted rules can work well but can take a lot of time to create and also restrict the way in which CSK can be extracted. This could lead to drastic changes in performance depending on the domain on which they are deployed.

"Open Information Generation: The Second Generation" [16] is a paper that was published in 2011 and attempts to improve upon TextRunner. In this paper two methods were introduced ReVerb & R2A2, representing the second generation of OIE methods and falling under the sub-category of rule-based OIE. As indicated by its name, ReVerb targets verbs and so every relational phrase is either a verb, a verb followed by a preposition, a verb followed by an adjective/noun or an adverb [6]. ReVerb differs from previous attempts, as it identifies the relation phrase by considering the entire phrase rather than word by word. Furthermore ReVerb considers multiple potential phrases, and then filters them by using it's own lexical constraint. The key aspect of R2A2 is that it introduces an argument identifier, named ArgLearner, to better extract arguments from relation phrases. ReVerb is good example of how restrictive rule-based OIE methods can be, as 65% of it's incorrect extractions were due to the argument-extraction heuristic failing [16]. This highlights how rule-based OIE methods can't adapt easily and thus are susceptible to making the same mistake over and over again.

**Hybrid OIE**

Hybrid OIE methods refers to methods that are combination of data-based and rule-based OIE concepts. This could be for several reasons, but it warrants it's own sub-category as certain methods wouldn't be acceptable in either the data-based or rule-based ones. Furthermore hybrid approaches show that it is possible to combine the two, as a way of working to the strengths of both.

"Commonsense Knowledge Extraction Using Concepts Properties" [8] is a paper published in 2011. It introduces a method which is based on the assumption that "concepts have properties which imply commonsense". The example they give to illustrate this idea is that "edible concepts can be found in a kitchen". The reason as to why this is a hybrid method is that it uses both rules that were automatically extracted and rules that were handcrafted. 83% of the rules were automatically extracted, but the human evaluation performed indicates that the handcrafted rules performed better, as they scored 4.26 (out of 5) on average as opposed to 3.81 (out of 5) [8]. This highlights how even though handcrafted rules perform better they are time consuming to create, and as such can still be useful to have automatically extracted rules.

### 3.5 Semi-Automatic: Human Validation

There are a lot less semi-automatic methods than automatic ones, but the majority of them fit under the 'Human Validation' umbrella. This category refers to methods where temporary results are automatically created and a human then has to validate them. There are variations when it comes to the power the human has, but as a base they can always accept or reject a statement. Certain methods however also allow humans to refine the statements, in order to enhance them.

"A semi-automated method for acquisition of commonsense and inferentialist knowledge"[32] is paper that was published in 2012 and introduces a semi-automatic method to collect CSK in Portuguese. The first step of this method, is for the user to enter a linguistic expression. From that there are two possibilities, the first one is that there already exists concepts for that expression and then they will be retrieved from the KB for the user to validate. Otherwise a new concept has to be created, and in that case the method consists of three separate steps. The first step performs a syntactic analysis of the input in order to define the structure of the noun phrase. In the second step heuristics are applied to pre-existing conceptual content in order generate more conceptual content. The third step is the validation step, in which the user has the ability to reject the baseline or include further common sense relations [32].

### 3.6 Semi-Automatic: Human Maintenance

This subsection of semi-automatic methods differs slightly from human validation, in that users can constantly oversee the repository rather than only when a formative result is extracted. For methods in this subsection users can go back to the repository any time and change statements that they've already amended. Whilst this is quite similar to the 'Human Validation' subsection, we felt splitting them up was justified as there aren't many semi-automatics published. As such it is a way of distinguishing between the methods that are published, even if it is only a nuance.

"A Semi-automatic Approach to Extracting Common Sense Knowledge from Knowledge Sources" [40] is a paper published in 2005 which introduces such a method. In this method the CSK is stored in a repository and provides an API for the user to modify it [3]. Having the ability to constantly refine the CSK means that the method is more scalable than other semi-automatic methods, as it makes the manual aspect less so of a bottleneck.

# 4 Effectiveness of the Methods

The previous section introduced the methods gathered and the categories they fall under. This section will be more focused on answering the core of the research question, which relates to the effectiveness of CSK acquisition methods. It will do so by first breaking down the research question, then discussing some metrics and finally discussing notable results.

The research question asks what the effectiveness of automatic and semi-automatic methods is in collecting CSK. Effectiveness is a broad term and thus it must be clarified in order to properly answer the question. For the purpose of this paper, effectiveness will be considered to be a measure of success. There are several ways to view success, but in the end it comes down to whether or not the methods have a real-world application. It is clear that CSK has a purpose in computer science, but it isn't clear how to go about gathering this CSK. Another key aspect to consider is whether or not the knowledge gathered is actually new knowledge and not just re-wording of previous knowledge.

## 4.1 Metrics

Metrics are a critical part of answering the question, as they are a way of measuring and quantifying success. There are a lot of different metrics, but in general they tend to fall under two sub-categories: automatic evaluation and human evaluation.

**Automatic Evaluation**
Automatic evaluation refers to metrics where a computer is able to automatically judge the performance often by using a scoring system. This isn't limited to such metrics however, as it also encompasses comparisons with pre-existing KBs. Similarly to how certain methods use already existing KBs to train, a part of that data can be used as a test set in order to judge the performance of a method.

An example of an automatic evaluation metric is the BLEU-2 scoring system [13]. It is the *go to* for automatic evaluations, as it is language independent and can be computed quickly. The problem with it however is that it tends to not correlate well with human judgement [13]. The BLEU-2 metric was used by COMET [10] in their evaluation, but since no other paper did so it can't be used for comparisons.

Another type of metric that is critical are novelty metrics. These metrics give an indication of how much of the knowledge is actually new. This is relevant for methods that require pre-existing knowledge on which to train, and such metrics give an indication of whether or not the method can actually be useful to generate new knowledge.

**Human Evaluation**
As previously mentioned CSK comes easily to humans, and as such a lot of evaluations rely upon humans to judge the output a method. There are variations in how the human evaluations are actually performed, certain are a binary decision where the human evaluator can decide whether or not they agree with the extracted knowledge. Others incorporate a scoring system, where they are to give each extraction a score between 1-5. Naturally these are all subjective and up to the interpretation of the individual, which is why most will have multiple human evaluators. Discrepancies in the way the human evaluation is performed can affect the results and is something to keep in mind.

From these human evaluations several papers use the same metrics. The first one and most common one is precision, which refers to the percentage of extractions returned which evaluators agreed with. Another metric used is recall, this refers to the percentage of all extractions that the method actually returned. From there it is possible to combine precision and recall by taking their harmonic mean creating the F-1 score [44].

This is a survey paper which doesn't actually perform tests in order to evaluate the performance of the methods and therefore it has to rely on the evaluations done in the papers gathered. This poses several problems, the most obvious one being that we're relying upon the creators of a particular method to evaluate their own work in a fair manner. It is natural that they would want to paint a good image of their methods but we have to trust their integrity. Another problem is that not all papers use the same process of evaluation and as such it can be hard to compare results. Certain papers do however compare their results with the ones from previous papers which facilitates things. If two specific papers don't share any metrics it is not possible to compare them, but they can be compared with other papers gathered that share the same metrics or they can be evaluated individually.

## 4.2 Notable Results

Going into depth about every method compiled for this paper is not possible and also not relevant. Discussing every result doesn't add much value, as simply re-stating results doesn't add any real value. It is more valuable to go into depth about a few representative result.

**Precision**
Precision is one of the most important metrics, as it indicates whether or not the knowledge is correct. Most of the time this is determined through a human evaluation, where judges are asked whether or not they agree with the extractions. For CSKBC methods it is also possible to do it by separating the data into training and test sets, to automatically evaluate the precision.

COMET [10] used two different CSK graphs, it used ATOMIC [37] and ConceptNet [20]. The human evaluation consisted of randomly selecting 100 events from the test set and generate the 10 most like inferences, and then presented them to 5 human evaluators. On the ATOMIC data it achieved a precision of 77.5% and 91.7% for ConceptNet. This

highlights how the underlying data can drastically affect the result, as the same method was used for both data sets and yet there is a significant difference in the outcome. Although these results are impressive they are still not quite as good as human performance, as the human precision was around 86% for ATOMIC.

MCN [2] also showed that it wasn't quite at the level of human performance in terms of precision, as they estimated the precision of a human to use as a comparison point. For the Reuters KB with 300 dimensional word embeddings the human estimate was a precision of 76.6%. In comparison, MCN achieved a precision of 62.8% with a standard deviation of 7.4, which is substantially worse.

**Recall**

Recall is a very important metric, as it gives an indication of how well a method uncovers the knowledge available. This is critical for CSK acquisition because computers need a huge amount of CSK for it to be worthwhile. For that reason it is not only important that a method produces reliable knowledge, but also that it produces sufficient amounts of knowledge. The issue is that recall requires some knowledge of the true amount of CSK available, which means it isn't always easy to calculate. A lot of papers tend to prioritize one over the other, as they see them as being a trade-off of one another [29].

ReVerb [16] highlights well how having low recall doesn't always have to be problematic. Since ReVerb is a very fast method that only has to go over the text once, it can be fed a lot more input. Therefore even though it will have a low recall, it can still create a substantial amount of CSK.

ASCENT [29] is different however, it aims to achieve both high precision and high recall. This was one of the goals of their method, to show that it is possible to target both recall and precision at the same time and not have to sacrifice one or the other. ASCENT was able to achieve a recall of around 75% whilst achieving a precision of 25.9%. This precision is deceptive however, as it represents the percentage of statements that received a rating of 5 out of 5 by human evaluators. Usually precision is measured as a binary metric, and if this had also been the case here it would've achieved much higher precision.

**F1-Score**

As previously mentioned, F1-Score is a combination of precision and recall. Some methods used this approach as a way of combining them, whilst others make a Precision-Recall curve and measure the area under that curve.

The evaluations of ReVerb [16] and DepOE [18] highlight how evaluations done similarly can still lead to different results. DepOE performed an evaluation in which they explicitly compared their method to ReVerb, as a way of highlighting how well their method performed. In that evaluation they found that ReVerb had an F1-Score of 0.44 [16], whilst in it's own evaluation is had an F1-Score of 0.61 [18]. This is a significant difference, which further emphasizes that results from metrics aren't the be all and end all.

**Novelty**

Novelty is especially interesting to consider for CSKBC methods, as they rely upon pre-existing knowledge. Metrics relating to novelty aren't interesting for OIE for example, as all the CSK they produce will be new. For CSKBC methods however it is a very telling metric, as it shows the true value of a method and its ability to improve coverage in a meaningful way.

COMET [10] discusses novelty by introducing two new metrics that automatically represent novelty. The first metric is the proportion of all tuples that are novel (% N/T *sro*). The second metric is the proportion of all tuples that have a novel object (% N/T *o*). Their method was able to have 59.25% of tuples not being present in the training set and more impressively 3.75% of nodes were new. The issue is that generations that are classified as being novel are actually just simplified forms of tuples that are present in the training set.

"Commonsense mining as knowledge base completion? A study on the impact of novelty" [21], as indicated by the title, also discusses novelty. This paper discuss the presence of an overlap between the training set and testing set, which leads to distorted results. For that purpose they introduce a novelty metric that relies upon word embeddings to calculate distance between statements. From that they were able to determine that there was an overlap between the training and testing set, which partially explains the discrepancy between KB completion and mining from wikipedia.

**Agreement Score**

Agreement score, also called Cohen's Kappa, is used to give an indication of the agreement between the human evaluators. It is more powerful than simply giving an agreement percentage as it accounts for the fact that an agreement can be by chance. A low agreement score means that the agreements are more random, whilst a higher agreement score means there is more agreement between the judges [33].

ReVerb [16] used this metric in their evaluation and got an agreement score of $\kappa = 0.68$, whilst the agreement percentage was 86%. It is an important metric, as it gives an indication of whether it is ambiguous or not. If there is a low agreement score can mean several things. It could be that the judges weren't given clear enough instructions. However if it isn't that, it could be that results are simply ambiguous, as CSK can be up to interpretation.

## 5 Discussion

The previous sections introduced a variety of categories with methods that work in completely different ways. These methods also serve different purposes which is shown in the way the methods work and their evaluations. Having considered methods that are more dated than others illustrates how much progress has already been made, and promises a bright future. Approaches have evolved over time, and as they keep evolving new categories could emerge.

As encouraging as the results are, they still have not reached human-level performance, especially in regards to

precision. Low precision in automatic methods can be problematic, since there is no human involved in the process to remove the wrong assertions. This means that the application that will be relying on the CSK created via that method, could be using faulty knowledge. Nevertheless automatic methods remain extremely valuable, as they're a lot more scalable and cheaper. The cost per statement of an automatic mining method is hundreds of folds less than that of a manually encoded one [41]. For that reason, and taking into account how much CSK is required the methods of CSK acquisition have to be fast and relatively cheap.

The different types of methods each have their strengths and weaknesses, making them each viable depending on the context. The mining methods don't rely upon any pre-existing KBs and as such they can be used on a new domain easily. Whilst they do have problems with precision and recall, they make up for it in their speed and scalability. CSKBC methods have higher precision and recall but they rely upon pre-existing KBs which often are manually created. It also has to be kept in mind that the evaluations for these methods overestimate their capacities, as re-wordings can appear as being novel knowledge. Nevertheless they are still able to increase coverage, which is very powerful as CSK graphs are huge and sparse [26].

Semi-automatic methods include humans in the process, which brings a level of quality control. Having a human oversee the results gives a level of confidence in the knowledge. Introducing a human in the process brings up some more problems, such as which humans should be doing that job and how many are needed. Another issue is the scalability, as humans will inevitably slow down the process and increase the costs. Depending on how much CSK is required, this could be problematic. As a result there are a lot less semi-automatic approaches than automatic ones, as scalability is one of the main problems to fix.

As of now, there is no *one size fits all* approach to CSK acquisition, and different approaches have different benefits. From the table in appendix A, it can be seen that there are significantly more automatic methods than semi-automatic ones. This doesn't mean however that there is no place for semi-automatic methods, but rather just that automatic methods were more suited to the problems being solved. It is about choosing the appropriate approach on a case by case basis. These approaches can also be used in combination with one another to complement each other.

## 6 Conclusions and Future Work

The field of CSK acquisition is still a relatively new one, but a lot of progress has already been made. As of now, none of the methods have reached human-levels of performance yet. Nevertheless recent methods, such as COMET [10], are sufficiently close to human performance and robust enough to be used to gather CSK for real world applications. This opens up new possibilities for AI, which is now tasked with figuring out the best way to use this CSK.

The approaches to the problem have evolved over time, and will continue to do so in the future. For that reason future work can also simply be going through the same process but later on and investigate the advances made. One of the shortcomings of this paper is that it didn't perform any tests itself, which made comparing different methods difficult. In the future it would be helpful to perform a controlled experiment on a variety of methods, to provide a basis from which to do comparisons. This would not only ensure that all the methods use the same metrics, but also have the same method of evaluation and data used for the evaluation. This would make it so that the comparisons would have some common ground, and aren't just speculative.

## 7 Responsible Research

The first thing to note is that this paper is a survey paper, and as such no experiments were performed as part of this project. The majority of the literature cited in this paper did however perform experiments.

It is possible that some of the literature surveyed selected may have tweaked their results or intentionally selected results that make their method look better. Although this is a possibility, we trust the integrity of the authors and the process to publish these papers. Furthermore several papers refer back to methods introduced by other papers and performed their own evaluations on those methods, providing another source of evaluation making it more trustworthy.

As this is a survey paper, the way in which the different literature is introduced and discussed, has to be in a fair and unbiased way. This paper should provide an unbiased overview of the field, and not try to push a hidden agenda. When comparisons between methods are made these have to be done in a fair manner and if there are factors that affect the comparison, they have to be disclosed. Furthermore any conclusions drawn have to be explained and have to be made clear. It has to be made clear what parts are opinions/deductions and which are coming from literature.

All the knowledge that is taken from the literature surveyed has to be appropriately cited. Naturally as this is a survey paper, a lot of the claims made will be from the literature, and thus the authors have to be credited appropriately. There also has to be a clear distinction between phrases which are quoted directly and those that are paraphrased.

Since all the work is cited, the process of this paper is highly reproducible. The methodology section also describes the process of creating the paper. As such reproducing the research should lead to similar conclusions.

## References


[1] Sakhawat Ali, Hamdy M. Mousa **and** M. Hussien. "A Review of Open Information Extraction Techniques". **in***IJCI. International Journal of Computers and Information*: (2019).

[2] Daniel Andrade **andothers**. "Leveraging knowledge bases for future prediction with memory comparison networks". **in***AI Communications*: 31 (6 2018), **pages** 465–483. ISSN: 09217126. DOI: 10.3233/AIC-170564.

[3] Christian Andrich, Leo Novosel **and** Bojan Hrnkas. *Common Sense Knowledge*. 2009. URL: http://www.uvm.nrw.de/opencontent.

[4] Gabor Angeli **and** Christopher Manning. "Philosophers are Mortal: Inferring the Truth of Unseen Facts". **in***Proceedings of the Seventeenth Conference on Computational Natural Language Learning*: Sofia, Bulgaria: Association for Computational Linguistics, **august** 2013, **pages** 133–142. URL: https://aclanthology.org/W13-3515.

[5] Michele Banko **andothers**. "Open Information Extraction from the Web". **in***Proceedings of the 20th International Joint Conference on Artifical Intelligence*: IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, **pages** 2670–2676.

[6] David Soares Batista **andothers**. *Large-Scale Semantic Relationship Extraction for Information Discovery*. 2016.

[7] Yonatan Bisk **andothers**. "PIQA: Reasoning about Physical Commonsense in Natural Language". **in***CoRR*: abs/1911.11641 (2019). arXiv: 1911.11641. URL: http://arxiv.org/abs/1911.11641.

[8] Eduardo Blanco, Hakki Cankaya **and** Dan Moldovan. *Commonsense Knowledge Extraction Using Concepts Properties*. 2011. URL: https://aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/view/2642.

[9] Katinka Böhm. "Semi-automatic engineering of topic ontologies from a common-sense knowledge graph". 2018. URL: https://repositum.tuwien.at/handle/20.500.12708/7715.

[10] Antoine Bosselut **andothers**. "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction". **in***CoRR*: abs/1906.05317 (2019). arXiv: 1906.05317. URL: http://arxiv.org/abs/1906.05317.

[11] Erik Cambria, Yunqing Xia **and** Amir Hussain. "Affective Common Sense Knowledge Acquisition for Sentiment Analysis". **in***Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*: Istanbul, Turkey: European Language Resources Association (ELRA), **may** 2012, **pages** 3580–3585. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/159_Paper.pdf.

[12] Erik Cambria **andothers**. *GECKA: Game Engine for Commonsense Knowledge Acquisition*. 2015. URL: http://rtw.ml.cmu.edu/rtw.

[13] Boxing Chen **and** Colin Cherry. *A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU*, **pages** 362–367. URL: http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

[14] Jacob Devlin **andothers**. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". **in***CoRR*: abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[15] Davis Ernest **and** Marcus Gary. "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence". **in***Communications of the ACM*: 58 (**september** 2015), **pages** 92–103. URL: https://cacm.acm.org/magazines/2015/9/191169-commonsense-reasoning-and-commonsense-knowledge-in-artificial-intelligence/fulltext?mobile=false.

[16] Oren Etzioni **andothers**. "Open Information Extraction: The Second Generation". **in***Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*: IJCAI'11. Barcelona, Catalonia, Spain: AAAI Press, 2011, **pages** 3–10. ISBN: 9781577355137.

[17] Joshua Feldman, Joe Davison **and** Alexander M. Rush. "Commonsense Knowledge Mining from Pretrained Models". **in***CoRR*: abs/1909.00505 (2019). arXiv: 1909.00505. URL: http://arxiv.org/abs/1909.00505.

[18] Pablo Gamallo, Marcos Garcia **and** Santiago Fernández-Lanza. "Dependency-Based Open Information Extraction". **in***Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*: Avignon, France: Association for Computational Linguistics, **april** 2012, **pages** 10–18. URL: https://aclanthology.org/W12-0702.

[19] Jonathan Gordon **and** Benjamin Van Durme. "Reporting bias and knowledge acquisition". **in***Proceedings of the 2013 workshop on Automated knowledge base construction*: (2013), **pages** 25–30.

[20] Catherine Havasi, Robert Speer **and** Jason B Alonso. *ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge*.

[21] Stanislaw Jastrzebski **andothers**. "Commonsense mining as knowledge base completion? A study on the impact of novelty". **in***CoRR*: abs/1804.09259 (2018). arXiv: 1804.09259. URL: http://arxiv.org/abs/1804.09259.

[22] Douglas Lenat. *CYC: A Large-Scale Investment in Knowledge Infrastructure Douglas B. Lenat*. 1995, **pages** 33–38.

[23] Xiang Li **andothers**. "Commonsense Knowledge Base Completion". **in***Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: Berlin, Germany: Association for Computational Linguistics, **august** 2016, **pages** 1445–1455. DOI: 10.18653/v1/P16-1137. URL: https://aclanthology.org/P16-1137.

[24] Chunhua Liu **and** Dong Yu. "BLCU-NLP at COIN-Shared Task1: Stagewise Fine-tuning BERT for Commonsense Inference in Everyday Narrations". **in***Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*: Hong Kong, China: Association for Computational Linguistics, **november** 2019, **pages** 99–103. DOI: 10.18653/v1/D19-6012. URL: https://aclanthology.org/D19-6012.

[25] Sap Maarten **andothers**. "Introductory Tutorial: Commonsense Reasoning for Natural Language Processing". **in**Association for Computational Linguistics (ACL): 2020, **pages** 27–33. ISBN: 9781948087803. DOI: 10.18653/v1/P17.

[26] Chaitanya Malaviya **andothers**. "Exploiting Structural and Semantic Context for Commonsense Knowledge Base Completion". **in***CoRR*: abs/1910.02915 (2019). arXiv: 1910.02915. URL: http://arxiv.org/abs/1910.02915.

[27] Mausam **andothers**. "Open Language Learning for Information Extraction". **in***Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: Jeju Island, Korea: Association for Computational Linguistics, **july** 2012, **pages** 523–534. URL: https://aclanthology.org/D12-1048.

[28] John Mccarthy. *PROGRAMS WITH COMMON SENSE*. 1959. URL: http://www.cs.cornell.edu/selman/cs672/readings/mccarthy-upd.pdf.

[29] Tuan-Phong Nguyen, Simon Razniewski **and** Gerhard Weikum. *Advanced Semantics for Commonsense Knowledge Extraction*. 2021.

[30] Simon Ostermann **andothers**. "Commonsense Inference in Natural Language Processing (COIN) - Shared Task Report". **in***Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*: Hong Kong, China: Association for Computational Linguistics, **november** 2019, **pages** 66–74. DOI: 10.18653/v1/D19-6007. URL: https://aclanthology.org/D19-6007.

[31] Lis Pereira **andothers**. "Adversarial Training for Commonsense Inference". **in***CoRR*: abs/2005.08156 (2020). arXiv: 2005.08156. URL: https://arxiv.org/abs/2005.08156.

[32] Vladia Pinheiro **andothers**. "A semi-automated method for acquisition of common-sense and inferentialist knowledge". **in***Journal of the Brazilian Computer Society*: 19 (1 **march** 2013), **pages** 75–87. ISSN: 16784804. DOI: 10.1007/s13173-012-0082-6.

[33] Kurtis Pykes. *Cohen's kappa*. **january** 2021. URL: https://towardsdatascience.com/cohens-kappa-9786ceceab58.

[34] Simon Razniewski, Niket Tandon **and** Aparna S. Varde. "Information to Wisdom: Commonsense Knowledge Extraction and Compilation". **in**Association for Computing Machinery, Inc: **august** 2021, **pages** 1143–1146. ISBN: 9781450382977. DOI: 10.1145/3437963.3441664.

[35] Christos Rodosthenous **and** Loizos Michael. "A hybrid approach to commonsense knowledge acquisition". **involume** 284: IOS Press, 2016, **pages** 111–122. ISBN: 9781614996811. DOI: 10.3233/978-1-61499-682-8-111.

[36] Itsumi Saito **andothers**. "Commonsense Knowledge Base Completion and Generation". **in***Proceedings of the 22nd Conference on Computational Natural Language Learning*: Brussels, Belgium: Association for Computational Linguistics, **october** 2018, **pages** 141–150. DOI: 10.18653/v1/K18-1014. URL: https://aclanthology.org/K18-1014.

[37] Maarten Sap **andothers**. "ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning". **in***CoRR*: abs/1811.00146 (2018). arXiv: 1811.00146. URL: http://arxiv.org/abs/1811.00146.

[38] Abhishek B. Sharma, Keith M. Goolsbey **and** David Schneider. "Disambiguation for Semi-Supervised Extraction of Complex Relations in Large Commonsense Knowledge Bases". **in**2020.

[39] Push Singh **andothers**. *Open Mind Common Sense: Knowledge Acquisition from the General Public*. 2002. URL: www.aaai.org.

[40] Veda C. Storey, Vijayan Sugumaran **and** Yihong Ding. "A Semi-automatic Approach to Extracting Common Sense Knowledge from Knowledge Sources". **in***NLDB*: 2005.

[41] Shane Storks, Qiaozi Gao **and** Joyce Y. Chai. "Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches". **in**(**april** 2019): CSK is implicit, comes naturally to humans¡br/¿Text mining is extimated to be hundreds of folds less than cost of manually encoding (Cyc). URL: http://arxiv.org/abs/1904.01172.

[42] Sangweon Suh, Harry Halpin **and** Ewan Klein. "Extracting Common Sense Knowledge from Wikipedia". **involume** 6: 2006. URL: www.geneontology.org.

[43] Niket Tandon, Gerard de Melo **and** Gerhard Weikum. "Deriving a Web-Scale Common Sense Fact Database." **injanuary** 2011.

[44] *What is the F1-score?* URL: https://www.educative.io/edpresso/what-is-the-f1-score.

[45] Hans Peter Willems. *ASTRID: Bootstrapping Commonsense Knowledge*. MindConstruct, **march** 2021. URL: https://www.mindconstruct.com.

[46] Chi-Hsin Yu **and** Hsin-Hsi Chen. "Commonsense Knowledge Mining from the Web". **in***AAAI*: 2010.

[47] Liang Jun Zang **andothers**. "A survey of commonsense knowledge acquisition". **in***Journal of Computer Science and Technology*: 28 (4 **july** 2013), **pages** 689–719. ISSN: 10009000. DOI: 10.1007/s11390-013-1369-6.

# A Overview of Methods Surveyed

| Overview of Methods Surveyed | | | | | |
|---|---|---|---|---|---|
| Branch | Sub-Category | Paper | Year | Method Name | Comments |
| Automatic | CSKBC | [10] | 2019 | COMET | Used both ATOMIC [37] and ConceptNet [20] KBs |
| | | [2] | 2018 | MCN | Gives an explanation for its predictions |
| | | [23] | 2016 | DNN AVG | Uses word averaging |
| | | | | DNN LSTM | Uses long short-term memory (LSTM) |
| | | | | Bilinear AVG | Uses word averaging |
| | | | | Bilinear LSTM | Uses long short-term memory (LSTM) |
| | | [26] | 2019 | Many Methods | All methods rely on ConvTransE decoder |
| | | [36] | 2018 | — | Learns CSKB Generation & Completion jointly |
| | | [38] | 2019 | —- | Ranks in terms of similarity to existing CSK |
| | | [4] | 2014 | — | Uses fact similarity to infer CSK |
| | Bootstrapping | [45] | 2021 | ASTRID | Inspired by how children learn |
| | | [43] | 2011 | — | Low recall considering amount of CSK on the Web |
| | Contextual Models | [31] | 2020 | ALICE | Only relies on the target dataset |
| | | [24] | 2019 | — | Fine tunes BERT [14] language model |
| | | [17] | 2019 | Coherency Rank | Uses bidirectional language model |
| | Rule-Based OIE | [16] | 2011 | ReVerb | Innovation is the relation phrase identifier |
| | | | | R2A2 | Better performance in evaluation than ReVerb |
| | | [29] | 2021 | ASCENT | Able to extract CSK from the Web |
| | | [18] | 2012 | DepOE | Used ReVerb as a performance benchmark |
| | | [42] | 2006 | — | Focuses on identifying generic statements |
| | Hybrid OIE | [8] | 2011 | — | Rules entered manually & automatically extracted |
| | Data-Based OIE | [27] | 2012 | OLLIE | Aims to improve upon ReVerb by using context |
| | | [5] | 2007 | TextRunner | Pioneer method in field of OIE |
| | | [46] | 2010 | — | Continuous learning of patterns |
| Semi-Automatic | Human Validation | [32] | 2013 | — | Made for Portuguese |
| | | [9] | 2018 | CN2TopicOnto | Extracts from ConceptNet, user defines axioms |
| | Human Maintenance | [40] | 2005 | — | Humans can Add/Modify/Delete CSK continuously |