The Weltmodell: A Data-Driven Commonsense Knowledge Base

Alan Akbik and Thilo Michael
Database Systems and Information Management Group (DIMA)
Technische Universität Berlin
Einsteinufer 17, 10587 Berlin, Germany
alan.akbik@tu-berlin.de, thilo.michael@gmail.com

Abstract

We present the WELTMODELL, a commonsense knowledge base that was automatically generated from aggregated dependency parse fragments gathered from over 3.5 million English language books. We leverage the magnitude and diversity of this dataset to arrive at close to ten million distinct N-ary commonsense facts using techniques from open-domain Information Extraction (IE). Furthermore, we compute a range of measures of association and distributional similarity on this data. We present the results of our efforts using a browsable web demonstrator and publicly release all generated data for use and discussion by the research community. In this paper, we give an overview of our knowledge acquisition method and representation model, and present our web demonstrator.

Keywords: Knowledge Acquisition, Open Information Extraction, Commonsense Reasoning

1. Introduction

Acquiring and representing structured, machine-readable general world knowledge has been a longstanding challenge in AI and NLP. Early approaches for the creation of commonsense knowledge were based on hand-coded knowledge by groups of expert logicians (Lenat, 1995). To overcome the inherent complexities in scaling up hand-built rule-systems, more recent approaches have investigated the use of crowd-sourcing (Singh et al., 2002) and the integration of numerous hand-built structured data sources (Speer and Havasi, 2012). A prominent example of this line of work is CONCEPTNET (Liu and Singh, 2004), a graph-based knowledge representation project in which nodes are concepts and edges relations that hold between these concepts.

We follow a different recent line of work which investigates the use of automated, data-driven methods to create commonsense knowledge bases (Schubert, 2002; Gordon et al., 2010; Gordon, 2010). We are motivated by ever-growing amounts of readily available natural language text (Halevy et al., 2009), an increasing maturity of Information Extraction (IE) technologies and the availability of scalable computing architectures (Dean and Ghemawat, 2008).

In particular, we are motivated by the availability of a dataset of aggregated dependency parse fragments gathered from over 3.5 million English language books (Goldberg and Orwant, 2013). By applying techniques from open-domain IE (Akbik and Löser, 2012) and distributional semantics (Akbik et al., 2012) to this large and diverse dataset, we investigate the potential of large scale data-driven approaches for automatically acquiring commonsense knowledge. We distinguish our work from related efforts (Speer and Havasi, 2012; Schubert, 2002) in that we model relations that hold between an arbitrary number of concepts (e.g. *N-ary relations*) and determine several notions of similarity and association which we include in our model. We make the results of our efforts publicly available for use and discussion by the research community.

We present the WELTMODELL, the commonsense knowledge base created with our data-driven approach, using a browsable web interface. Typical commonsense information that users may query in the WELTMODELL includes:

- Facts that pertain to a given concept, along with co-occurrence counts and mutual information values. For instance, a user might query for the concept "coffee" (see Table 1). Example facts include unary facts ("***coffee*** *may smell good*"), binary facts ("***someone*** *may drink* ***coffee***") and ternary facts ("***someone*** *may have* ***coffee*** *for* ***breakfast***").
- Concepts that may fill a slot in a given statement. For example, given the facts found for "coffee" in Table 1, users may ask: "*What are other things that someone may have for breakfast?*", "*What other things may smell good?*", or "*What things may a waiter bring?*". See Table 2 for example results for such queries.
- Statements with high *applicative similarity* for a given statement. For instance, a user might ask "*What statements hold for things that a waiter may bring?*" (they may be put or placed on a table). We discuss such notions of similarity in section 2.4.

In the following, we give an overview of our knowledge acquisition methodology and knowledge representation model. We also give a brief overview of our web UI.

FACT	NPMI	COUNT
[someone] may sip [**coffee**]	0.65	6946
[someone] may drink [**coffee**]	0.56	11206
[someone] may pour [**coffee**]	0.52	2658
[**coffee**] may smell good	0.49	219
[someone] may have [**coffee**] for [breakfast]	0.44	102
[waiter] may bring [**coffee**]	0.41	140

Table 1: Facts that pertain to the concept "**coffee**", ordered by normalized pointwise mutual information (NPMI). Facts have different numbers of slots for concepts which are indicated by brackets. Thus, "[someone] may have [**coffee**] for [breakfast]", for example, is a ternary statement.

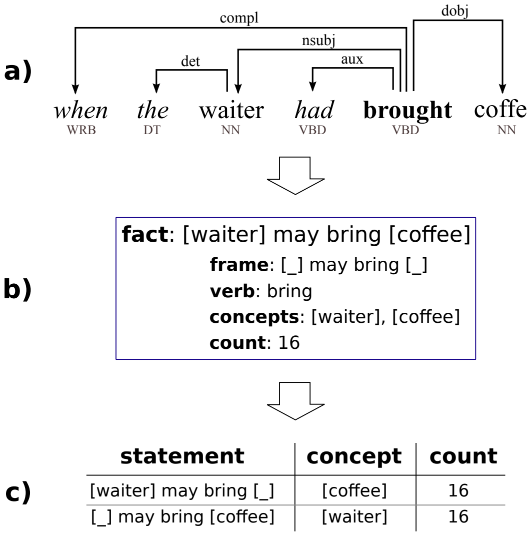
a) [someone] may sip **[_]**		**b)** [someone] may have **[_]** for [breakfast]		**c)** [waiter] may bring **[_]**		**d)** **[_]** may smell good	
CONCEPT	NPMI	CONCEPT	NPMI	CONCEPT	NPMI	CONCEPT	NPMI
coffee	0.65	mackerel	0.62	check	0.48	gunpowder	0.49
brew	0.58	sausage	0.58	coffee	0.41	coffee	0.49
tea	0.57	cocoa	0.56	drink	0.41	cooking	0.36
drink	0.5	oatmeal	0.51	salad	0.4	soup	0.35
champagne	0.49	omelette	0.49	tray	0.38	food	0.34
wine	0.48	coffee	0.44	menu	0.36	air	0.32

Table 2: Typical concepts that are observed for four example statements of different arity. Each statement has exactly one free slot, indicated as "[_]". In example **b)**, a ternary statement, the concepts observed that may fill the free slot are different items of food that one may have for breakfast. Example **d)** is a unary statement that is observed for concepts that may smell good.

2. Knowledge Acquisition

2.1. Data set

We base our knowledge acquisition effort on the Syntactic N-Grams dataset (Goldberg and Orwant, 2013). It was extracted from corpus of 3.5 million digitized English Google Books (Michel et al., 2011) and includes ca. 10 billion distinct items. The dataset contains *syntactic n-grams*, rooted dependency tree fragments from parsed English text. N-grams may consist of up to 5 arcs in a dependency tree between *content-words* (word types judged to be semantically important such as nouns and verbs), plus any number of arcs to *functional-markers*, word types such as determiners and auxiliary verbs. At time of writing, it is the largest openly available corpus of its kind. An example of a syntactic n-gram that was seen 16 times is given in Figure 1a).



fact: [waiter] may bring [coffee]
frame: [_] may bring [_]
verb: bring
concepts: [waiter], [coffee]
count: 16

statement	concept	count
[waiter] may bring [_]	[coffee]	16
[_] may bring [coffee]	[waiter]	16

Figure 1: Extraction process for an example n-gram. The n-gram is displayed in *a)*. The extraction method identifies the head verb (highlighted bold) and dismisses all nodes that are not of interest (highlighted in italics). The extracted fact with its components (frame, verb, concepts and count) is listed in *b)*. The collapsed statement-concept representation for this fact is given in *c)*.

2.2. Fact Extraction

Our method expects as input a dependency tree fragment that contains a verb and all of its arguments. We apply an open-domain Information Extraction method modeled on the KRAKEN system described in (Akbik and Löser, 2012): Using a simple rule-set over typed dependencies, we collect subjects, particles, negations, passive subjects, direct and prepositional objects of the verb, and determine whether the dependency tree fragment contains a passive construction, and whether it is negated. Personal pronouns and certain nouns (such as "anybody" and "everyone") are replaced with the placeholder "someone". Some vague words such as "that" and "everything" are generalized to "something". Verbs are lemmatized in active, but not in passive, constructions.

We then place the collected arguments in the following order: Subject, negation, main verb, particles, direct objects and prepositional objects. All other links, such as auxiliary verbs and complements are disregarded. So, the input fragment "*when the waiter had brought coffee*" yields the fact "[waiter] may bring [coffee]". Refer to Figure 1 a) and b) for an illustration of this step.

Representation. This yields a large resource of N-ary facts. Facts consist of a *frame* with slots for each noun concept. So, in the fact "[waiter] may bring [coffee]", the frame is "[_] may bring [_]" and the concepts are "waiter" and "coffee". Frames are always based on a *verb*, so that for each verb a number of frames with different numbers of slots and prepositions exist. The verb "bring", for instance, has frames such as "[_] may bring [_] before [_] " and "[_] may bring down [_]". Aggregated frequency *counts* from the Syntactic N-Grams dataset are also stored for each fact. An illustration of a fact and its components is given in Figure 1 b).

Finally, a *statement* is a fact in which exactly one slot is empty, such as "[waiter] may bring [_]". Statements pertain to concepts that may fill the empty slot. This concept-statement pair representation is important to our knowledge base, because it allows us to define filters and calculate several measures of similarity and association.

2.3. Pointwise Mutual Information.

Many frequently observed statements are very general and are seen with very large numbers of concepts. An example

FUNCTIONAL SIMILARITY		
Concept 1	Concept 2	Cosine
coffee	**tea**	0.54
coffee	brew	0.54
belief	notion	0.53
belief	assumption	0.46
belief	idea	0.45

APPLICATIVE SIMILARITY		
Statement 1	Statement 2	Cosine
[_] may cross [face]	[_] may pass over [face]	0.25
[_] may hang in [air]	[_] may fill [air]	0.62
[_] may hang in [air]	[_] may fill [room]	0.66
[_] may hang in [air]	[someone] may exhale [_]	0.67
[_] may cure [disease]	**[someone] may be administered [_]**	0.86

Table 3: Examples for functional and applicative similarities. The concepts "coffee" and "tea" (highlighted bold) share a high functional similarity, because statements that hold for "coffee" often also hold for "tea" (such as "[someone] may drink [_]"). The statements "[_] may cure [disease]" and "[someone] may be administered [_]" because many things that cure diseases may also be administered to someone.

is "[someone] may see [_]" which is observed for nearly any noun. In order to measure the strength of the association between a statement and a concept that may fill the statement's free slot, we compute the mutual information of each concept-statement pair. We use the *pointwise mutual information* (PMI) - as well as its normalized variant (NPMI) - and calculate the discrepancy between the probability of a concept-statement coincidence given their joint distribution and their individual distributions, assuming independence. Equation 1 shows the formula for the NPMI, where $p(s)$ is the probability of a statement and $p(c)$ the probability of a concept.

$$npmi() = log\frac{\frac{p(s,c)}{p(s)*p(c)}}{-log[p(s,c)]} \qquad (1)$$

We also compute the so-called Lexicographer's Mututal Information (LMI) by multiplying the NPMI with the concept-statement co-occurrence counts. This offsets the tendency of the PMI to rate rare events to highly. We make all three statistics available, as we feel that each may be used to find significant, or *typical*, statements for a given concept and vice versa.

2.4. Functional and Applicative Similarity

We compute measures of similarity for all pairs of concepts and all pairs of statements.

Functional similarity. Following (Turney, 2012), we define the *functional similarity* of two concepts to be the degree to which two concepts are observed with the same statements. The concepts "coffee" and "tea" for example are functionally similar, because they share a large set of statements, such as "[someone] may drink [_]", "[someone] may sip [_]" and "[someone] may pour [_]". See Table 3 for examples.

Applicative similarity. We define two statements to have a high *applicative similarity* if they are observed with similar nouns. For example, the statements "[_] may cure [disease]" and "[someone] may be administered [_]" have a certain applicative similarity because they are observed with nouns such as "medicine", "dose" and "draught". A verbalization of this example would be "*things that cure diseases may often be administered to someone*". See Table 3 for examples.

We measure the similarity of two concepts or statements using the *cosine distance* (Bullinaria and Levy, 2007) over shared attributes in a vector space presentation. The closer the cosine distance is to 0, the more similar two concepts or statements are. We describe the process of calculating the cosine distance in greater detail in (Akbik et al., 2012).

2.5. Results

We apply our algorithm to the dataset and find 852,387,621 facts, of which 9,544,862 are distinct. The knowledge base spans 933,997 distinct frames, 2,993,678 distinct concepts and 6,155,115 distinct statements.

3. Demonstration

For demonstration and discussion purposes, we make available a browsable web interface to the WELTMODELL at http://www.textmining.tu-berlin.de/weltmodell, where users can freely browse concepts, statements, frames and verbs. The two most important views are the concept and the statement views:

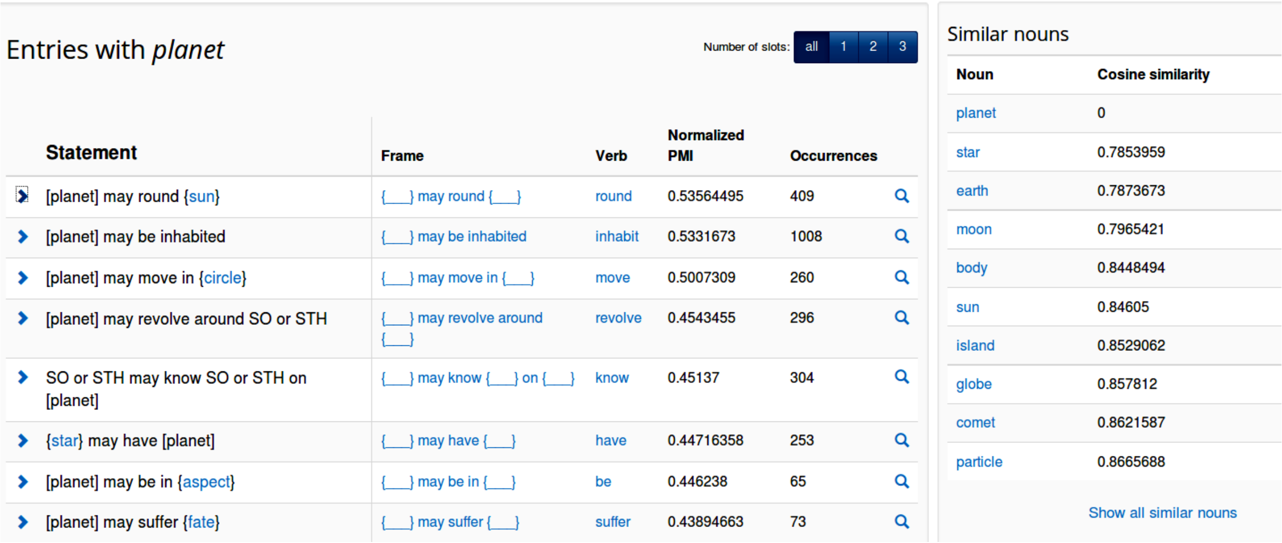
Concept view The concept view shows associated statements and functionally similar concepts for a given concept. Users may filter statements according to their arity. See Figure 2 for an example. A click on the arrow next to a statement leads to the statement view.

Statement view The statement view shows associated concepts and applicatively similar concepts for a given statement. Here, users can execute a single-link hierarchical agglomerative clustering (HAC) by clicking the "group" button. Associated concepts are then grouped according to their functional similarity. We show an example of this view and the visualized clustering in Figure 3.

Users can compare the different mutual information metrics and inspect for each fact a sample of syntactic n-grams in which it was found.

4. Outlook

Present work focuses on expanding the range of our knowledge acquisition efforts to non-verb constructions such as noun phrases, and finding causal relationships between verbs (Rink et al., 2010; Gordon et al., 2011). We intend to increase the data quality by continuously refining our extractors and integrating external repositories such as WORDNET (Miller, 1995) and OPEN MIND COMMON SENSE (Singh et al., 2002). We also plan to add more visualization options as we add new types of commonsense



Entries with *planet*

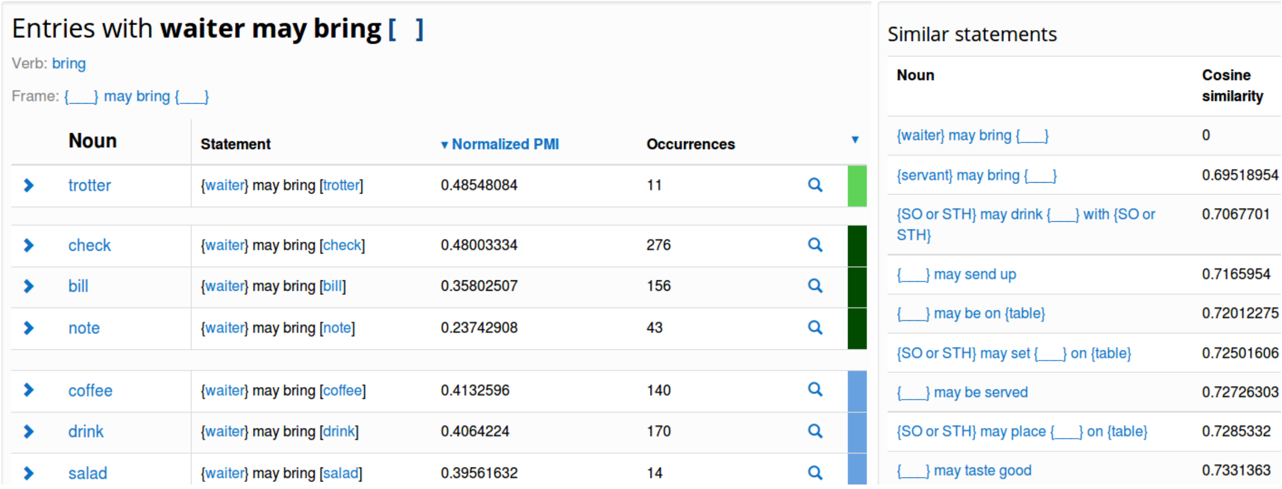
	Statement	Frame	Verb	Normalized PMI	Occurrences	
›	[planet] may round {sun}	{___} may round {___}	round	0.53564495	409	🔍
›	[planet] may be inhabited	{___} may be inhabited	inhabit	0.5331673	1008	🔍
›	[planet] may move in {circle}	{___} may move in {___}	move	0.5007309	260	🔍
›	[planet] may revolve around SO or STH	{___} may revolve around {___}	revolve	0.4543455	296	🔍
›	SO or STH may know SO or STH on [planet]	{___} may know {___} on {___}	know	0.45137	304	🔍
›	{star} may have [planet]	{___} may have {___}	have	0.44716358	253	🔍
›	[planet] may be in {aspect}	{___} may be in {___}	be	0.446238	65	🔍
›	[planet] may suffer {fate}	{___} may suffer {___}	suffer	0.43894663	73	🔍

Similar nouns

Noun	Cosine similarity
planet	0
star	0.7853959
earth	0.7873673
moon	0.7965421
body	0.8448494
sun	0.84605
island	0.8529062
globe	0.857812
comet	0.8621587
particle	0.8665688

Show all similar nouns

Figure 2: Information on the concept "planet". On the left side, some associated statements are listed. On the right side, concepts with high functional similarity are listed, such as "star", "earth" and "moon".



Entries with **waiter may bring [ ]**

Verb: bring

Frame: {___} may bring {___}

	Noun	Statement	▾ Normalized PMI	Occurrences		▾
›	trotter	{waiter} may bring [trotter]	0.48548084	11	🔍	🟩
›	check	{waiter} may bring [check]	0.48003334	276	🔍	🟩
›	bill	{waiter} may bring [bill]	0.35802507	156	🔍	🟩
›	note	{waiter} may bring [note]	0.23742908	43	🔍	🟩
›	coffee	{waiter} may bring [coffee]	0.4132596	140	🔍	🟦
›	drink	{waiter} may bring [drink]	0.4064224	170	🔍	🟦
›	salad	{waiter} may bring [salad]	0.39561632	14	🔍	🟦

Similar statements

Noun	Cosine similarity
{waiter} may bring {___}	0
{servant} may bring {___}	0.69518954
{SO or STH} may drink {___} with {SO or STH}	0.7067701
{___} may send up	0.7165954
{___} may be on {table}	0.72012275
{SO or STH} may set {___} on {table}	0.72501606
{___} may be served	0.72726303
{SO or STH} may place {___} on {table}	0.7285332
{___} may taste good	0.7331363

Figure 3: Statement view for "[waiter] may bring [_]", i.e. all things that a waiter may bring. Concepts are grouped according to functional similarity, so one functionally similar group of concepts a waiter may bring is "check", "bill" and "note". Applicatively similar statements are listed on the right. The statement "[someone] may place [_] on [table]" is similar. This reads as "*things that a waiter may bring may also often be placed on a table*".

information. We will also investigate in how far we can encourage and automatically integrate user feedback into our knowledge base.

Disambiguation. One principal problem that we are addressing is the widespread ambiguity of concepts (as well as statements to a lesser extend). Examples for ambiguous concepts are "plane" (airplane vs. field) and "change" (coin vs. modification). Our hope is to separate the meanings either using unsupervised methods, for example through clustering the statements for each concept, or by disambiguating concepts to WordNet. Initial investigations of both supervised and unsupervised approaches point to potential, however no results of sufficient quality for inclusion into the publicly browsable dataset have yet been produced.

Incompleteness. Another challenge that we are investigating concerns incompleteness; even with the very large dataset used in our experiments, not all possible facts are observed. An example of this is the statement "[_] may have [wing]", which is observed for many different birds, but not the concept "owl". However, the WELTMODELL contains enough information to infer such statements, e.g. we know that an owl is a bird and that it can fly. We also observe that birds that can fly often have wings. Currently we are investigating matrix completion methods as a general means to infer new information (Riedel et al., 2013).

Common sense reasoning. In order to gather further requirements for data-driven commonsense knowledge acquisition, we intend to address commonsense reasoning tasks such as the Choice of Plausible Alternatives (COPA) task described in (Roemmele et al., 2011). Our goal is to create methods that not only heuristically solve such tasks but also "explain" their choice as a series of human-readable commonsense reasoning steps.

5. Acknowledgements

We would like to thank our colleague Larysa Visengeriyeva for her help and ideas in working with the syntactic n-grams. Alan Akbik and Thilo Michael received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT-2009-4-1 270137 'Scalable Preservation Environments' (SCAPE).

6. References

Akbik, A. and Löser, A. (2012). Kraken: N-ary facts in open information extraction. In *AKBC-WEKEX*, pages 52–56. Association for Computational Linguistics.

Akbik, A., Visengeriyeva, L., Herger, P., Hemsen, H., and Löser, A. (2012). Unsupervised discovery of relations and discriminative extraction patterns. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Bullinaria, J. and Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of english books.

Gordon, J., Van Durme, B., and Schubert, L. K. (2010). Learning from the web: Extracting general world knowledge from noisy text. In *Collaboratively-Built Knowledge Sources and AI*.

Gordon, A. S., Bejan, C. A., and Sagae, K. (2011). Commonsense causal reasoning using millions of personal stories. In *AAAI*.

Gordon, A. S. (2010). Mining commonsense knowledge from personal stories in internet weblogs. *Automated Knowledge Base Construction*, page 8.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Liu, H. and Singh, P. (2004). Conceptnet - a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas.

Rink, B., Bejan, C. A., and Harabagiu, S. M. (2010). Learning textual graph patterns to detect causal event relations. In *FLAIRS Conference*.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Schubert, L. K. (2002). Can we derive general world knowledge from texts? In *Proceedings of the second international conference on Human Language Technology Research*, pages 94–97. Morgan Kaufmann Publishers Inc.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *J. Artif. Intell. Res.(JAIR)*, 44:533–585.