Identifying Underlying Commonsense Knowledge in Definitions

Jansen Orfan, James Allen
Department of Computer Science, University of Rochester
Rochester, New York 14627 USA

Abstract

We present a framework that learns commonsense temporal knowledge from word definitions. Our work differs from existing systems in both the way definitions are axiomatized and the way knowledge is inferred from those axioms. First, we go beyond axiomatizing just the literal interpretation of a definition by considering the underlying subtext and assumptions a reader has to make to understand a definition. Secondly, we cluster the concept axioms into small event theories that we use to predict the co-occurrence of concepts in simple scenarios. These predictions allow us to identify knowledge derived from the complex interactions among several definitions that would otherwise be ignored. We show that this framework can derive temporal knowledge across several different concept domains. Results are compared to human judgment and demonstrate the effect several features have on evaluation scores.

Introduction

Common-sense temporal knowledge is one of the several types of knowledge needed for general purpose natural language understanding. Such knowledge helps infer a more complete narrative from a statement. For instance, from "Kim was kept up until 2 a.m. by the music." and the definition of **keep_up$_{v5}$**, *"prevent from sleeping"*, we can infer that Kim went to sleep around 2 a.m., she intended to go to sleep before then, and the music probably stopped around that time. Unfortunately, it is not feasible to hand produce a large enough lexical knowledge base (KB) for this task. Instead, automated methods are employed to extract knowledge from natural language.

Systems that focus primarily on facts about instances (e.g. *Nice is in France*) as well as concept subsumption and mereology, like NELL (Carlson et al. 2010), can take advantage of large corpora because such knowledge is often encoded in surface features like syntax patterns and selectional preferences. However, finding concept relations (e.g. *sleeping precedes waking up*) in most forms of natural language is more difficult. Concept relations are assumed to be so well known that most language adhering to Grice's maxims (Grice 1975), particularly that communication should "be brief", would omit any useful indication of their nature. One would not normally say "Kim woke up and is no longer sleeping" because the speaker would assume the audience understands the connection between waking up and sleeping. There is, however, a type of natural language that does make concept relations more explicit: concept glosses (short definitions) found in dictionaries assume the audience is unfamiliar with the concept and will discuss its relations to other concepts more explicitly than most other language.

Approaches that axiomatize glosses - like (Harabagiu, Miller, and Moldovan 1999) and more recently (Clark et al. 2008), (Allen et al. 2013), and (Kim and Schubert 2016) - differ significantly from corpora-based ones. Most apparently, the input for gloss-based systems is smaller but the information density is much higher - a good gloss directly states a concept's most important qualities and the qualities of other closely related concepts. Although gloss-based approaches could be applied to any dictionary, WordNet (WN) (Miller et al. 1990) is preferred over most other sources because many of its glosses have been sense tagged and WN provides additional concept knowledge among those senses (e.g. hypernym and antonym relations). Furthermore, WN senses are widely used in other NLP applications which allows for better integration with other projects.

The more recent gloss approaches referenced above work similarly, but chiefly differ in their representation of the knowledge extracted. Each parses WN glosses into logical forms (LF) consisting of word senses and thematic roles. They assert that each concept entails its LF (i.e. $Concept \rightarrow LF_{Gloss}$). If the WN gloss for **waken$_{v2}$** is, *"stop$_{v1}$ sleeping$_{n1}$"* then the resulting LF would be similar to Figure 1.

With general event axioms and logical inference one could learn relationships like event pre- and post-conditions and entailment. Figure 2 shows a small portion of knowledge we can extract just from LFs and WN's concept relations. Although we can extract a great deal of knowledge, it

$$waken_{v2}(e1) \wedge agent(e1, x) \rightarrow$$
$$stop_{v1}(e2) \wedge sleeping_{n1}(e3) \wedge agent(e2, x) \wedge$$
$$effect(e2, e3) \wedge agent(e3, x).$$

Figure 1: Axiom from **awaken$_{v2}$** - *"stop$_{v1}$ sleeping$_{n1}$"*

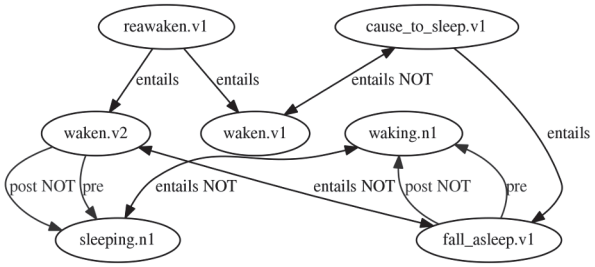

Figure 2: Some of the axioms generated directly from the definitions of concepts related to **waken$_{v1}$**.

remains sparsely connected. For instance we would expect **waken$_{v1}$**,*"cause to become awake or conscious"* to have most, if not all, of the relationships that **waken$_{v2}$** has.

We only have one gloss per concept from which to derive knowledge and we, again, come up against the tendency to "be brief". Human readers derive knowledge not only from the literal meaning of a gloss (like its LF) but also interpret the subtext - the intended meaning of the gloss - and understand it in the context of related concepts. In previous work (Orfan and Allen 2015) we explored areas of subtext that current methods do not exploit and logical inference alone cannot handle. Using hand-axiomatized glosses related to **sleep$_{n1}$** and a general theory about events and time, we showed that including such underlying information allows us to extract more knowledge from definitions.

In this paper we expand on the ideas in (Orfan and Allen 2015) by increasing the flexibility of the temporal theory, automating the axiomatization method, and showing that these methods can be applied to concept domains beyond just **sleep$_{v1}$**. We present a framework that extracts temporal relations from both the literal and sub-textual information found in WN definitions. Axioms are extracted based on both a gloss' LF and its underlying meaning. Those axioms are then combined using a simple temporal model to infer new temporal relations that would otherwise be overlooked. We evaluate our results based on human judgment and show that this framework extracts more relationships than current methods at a comparable level of precision.

System Overview

Our framework can be broken up into several distinct steps (outlined in Figure 3). First we choose the concept we wish to learn about ("Seed Concept"). From that concept we build a set of neighboring concepts (like those in Figure 2) about events and states close to the seed via LF graphs and WN's concept relations (e.g. hypernym, antonym, and derivationally-related links). For each concept in the set we convert their gloss' LF (both the literal interpretation and subtext) and its concept relations to temporal axioms ("Definition to Axioms"). The axioms are combined with general temporal axioms ("Temporal Theory") to produce a small theory ("Micro-Theory") that characterizes the seed sense by describing how the concepts related to it interact.

Using these micro-theories we infer ("Inference") the state of the world given several basic premises about the seed

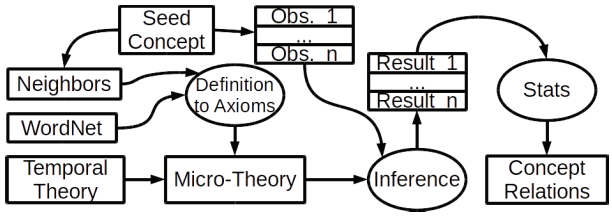

Figure 3: Diagram showing the flow of input through different processes to eventually yield concept relations. Rectangles represent data and circles represent processes.

sense ("Obs.1") like, "Something was sleeping and now is not sleeping." These inferences give insight into how events and states interact according to their definitions. We do this for several similar premises and infer temporal knowledge from co-occurrence ("Stats"). For instance, if every time we assert that someone is **asleep$_{adj1}$** the micro-theory predicts they are also **sleeping$_{n1}$**, then we would infer that **asleep$_{n1}$** entails **sleeping$_{n1}$**

In the next section we present the temporal theory we use to axiomatize concepts. Then we show how definitions are axiomatized and combined to drive inference. We then describe how micro-theories are used to make predictions and how those are used to infer common sense temporal knowledge.

Interval Temporal Logic

We represent glosses and subtext in a simplified derivative of interval temporal logic (ITL). ITL was developed as a representation in (Allen 1984) and further expanded to support plan reasoning and event prediction in (Allen and Ferguson 1994). ITL was constructed with the purpose of representing and reasoning about the complexities involved in several temporal aspects of events and states in natural languages. However, we have found that definitions predominately rely on only a few verb aspects: stative, progressive, inceptive, terminative, and preventative. For the sake of speeding up event prediction, we created a simplified implementation of ITL to omit some of the complex temporal relations that we do not require.

Our implementation represents time as a finite number of discrete steps that cannot overlap; time steps can be consecutive. e.g. $meets(t1, t2)$ means that $t1$ is directly before $t2$. For convenience, assume that $meets(t, t+1)$ is always true. We represent two types of lexical concepts, **states** and **events**, defined below. In practice we distinguish between these two concept types based primarily on where they fall in WN's hypernym hierarchy.

- **State** (s) – properties/states that objects may have/be in during a given time step (*e.g.* Awake, Snore)
- **Event** (e) – processes that affect states when they start/stop (*e.g.* Waken, Fall Asleep, Kill)
- **Time** (t) – a discreet span of time; a time step

In our model, states are said to $hold$ at a certain time. Events can $occur$ over one or many time steps and are said

$hold(sleeping_{n1}, 1)$ $\quad \neg hold(sleeping_{n1}, 2)$
$start(waken_{v2}, 1)$ $\quad stop(waken_{v2}, 2)$
$occur(waken_{v2}, 1)$ $\quad \neg occur(waken_{v2}, 2)$

Figure 4: Representation of a scenario where something is sleeping to begin with, then wakes and is no longer sleeping.

to $start$ at the first time step they begin to $occur$ and $stop$ at the first time step they no longer $occur$. Figure 4 provides an example describing a scenario with this representation.

- $meets$(**Time**,**Time**) - two time steps are sequential
- $hold$(**State**, **Time**) - a state is true at a certain time.
- $occur$(**Event**,**Time**) - an event is occurring at a certain time.
- $start$(**Event**,**Time**) - an event starts occurring at a certain time, is occurring, and its preconditions are true
- $stop$(**Event**,**Time**) - an event stops occurring at a certain time, is not occurring, and its postconditions are true

We define a set of general axioms to reason about events. For the sake of space, we will not discuss axioms defining the semantics of the predicates already described above. One departure we make from (Allen and Ferguson 1994) is our handling of explanation closure axioms. Such axioms dictate that when a state, s, changes then some event that is known to change s must have caused it. However, given our task, we have to assume that our knowledge of events is incomplete and so we cannot be sure what does and does not change s - all we know is that *some* event caused it to change. We use a weakened version of the closure axioms which essentially states that if s changes states between two time steps, then some event has either started in the first step or stopped in the second. Formally: $\forall t.hold(s,t) \neq hold(s,t+1) \rightarrow \exists e(start(e,t) \vee stop(e,t+1))$

Axioms from Glosses

Now that we have a logic to represent definitions we can move on to axiomatizing them. In this paper we focus on three types of temporal relationships between states and events: entailment, precondition, and postcondition. Their formal semantics are shown in Figure 5. We will start by describing how knowledge is extracted from literal interpretations of glosses and then describe how their underlying meanings are added.

Like similar approaches we begin by parsing WN glosses. These parses are generated using the TRIPS parser (Allen 2014), which was previously used in (Allen et al. 2013). TRIPS is a semantic parser with several features that make it well suited for understanding definitions. Importantly, it

s ENTAILS s' $\quad \forall t.hold(s,t) \rightarrow hold(s',t)$
e ENTAILS e' $\quad \forall t.occur(e,t) \rightarrow occur(e',t)$
e PRE s $\quad \forall t.start(e,t) \rightarrow hold(s,t)$
e POST s $\quad \forall t.stop(e,t) \rightarrow hold(s,t)$

Figure 5: Semantics of temporal relationships.

can tag words with WN senses in cases where a sense tag is not provided. Furthermore, words are linked to a linguistically motivated ontology that allows us to easily define general rules for extracting knowledge. Other works use similar methods to extract semantic relationships from LFs. For instance, (Clark et al. 2008) includes a set of generic axioms that connect key WN senses (like **become$_{\mathbf{v1}}$**) to semantic relationships like CHANGETO and CHANGEFROM.

We created about 30 LF patterns that extract temporal relationships based on the structure of the LF and the ontology. An example pattern is shown below:

$stop(x) \wedge effect(x,y) \Rightarrow x$ PRE y and x POST $\neg y$
if x is subsumed by the concept **stop** and has an *effect* relation to y then y is a precondition of x and $\neg y$ is a postcondition of x

In the case of the LF in Figure 1, we would extract the following two relationships: $waken_{v1}$ POST $\neg sleeping_{n1}$ and $waken_{v1}$ PRE $sleeping_{n1}$. Similarly for **stay_up$_{\mathbf{v1}}$**, *"not go_to_bed$_{vl}$"* we would extract: $stay\_up_{v1}$ ENTAILS $\neg go\_to\_bed_{v1}$.

We take advantage of the concept relationships that WN provides, namely hypernym and antonym links. The relation, x HYPERNYM y is treated like x ENTAILS y and x ANTONYM y is treated like x ENTAILS $\neg y$.

We can extract a good amount of knowledge from just the literal interpretation of glosses and logical entailment, as (Clark et al. 2008), (Allen et al. 2013), and (Kim and Schubert 2016) do successfully. But the extracted knowledge is very sparse and disconnected. They only have one gloss from which to learn and each is only a few words long - there simply is not enough space to properly characterize a concept. If we want to extract more knowledge then we have to look beyond the literal interpretation of a gloss.

Adding Subtext

Below we have listed four intuitions that give insight on how we can extract more information from a definition beyond just axiomatizing the LF.

1. Negation in language can mean more than just logical negation. The definition **stay_up$_{\mathbf{v1}}$**, *"not go_to_bed$_{vl}$"* does not simply mean $stay\_up_{v1}$ ENTAILS $\neg go\_to\_bed_{v1}$. The reader knows that **stay_up$_{\mathbf{v1}}$** is in the same domain as **go_to_bed$_{\mathbf{v1}}$** and entails at least some of **go_to_bed$_{\mathbf{v1}}$**'s precondtions and precludes some of its postconditions. In this case *"not go_to_bed$_{vl}$"* implies that you are able to go to bed - i.e. you are awake.
2. Glosses are meant to closely approximate the concept they define, but a gloss cannot typically replace its concept without losing meaning. While $Concept \rightarrow LF_{Gloss}$ is true, limited space and other pragmatic considerations mean that in most cases we should not also introduce axioms of the form, $LF_{Gloss} \rightarrow Concept$ (Stock 1988). But a gloss does provide the reader strong evidence for its concept. Unfortunately this is something that is difficult to manage using only logical entailment. Consider, **sleeping$_{\mathbf{n1}}$**, *"the state of being asleep"* and **snore$_{\mathbf{v1}}$**, *"breathe noisily during one's sleep"*. **sleeping$_{\mathbf{n1}}$**'s gloss is

an excellent replacement for **sleeping$_{\mathbf{n1}}$** but the gloss for **snore$_{\mathbf{v1}}$** misses some of the subtleties that distinguishes snoring from loud breathing. We cannot expect a dictionary to explain the physiological causes of snoring so we have to settle for a less exact description.

3. Derivationally-related links very often point to different linguistic realization of a state or event. For instance **sleep$_{\mathbf{n1}}$**,*"a natural and periodic state of rest during which consciousness of the world is suspended"* has links to **sleep$_{\mathbf{v1}}$**, *"be asleep"* and **sleepy$_{\mathbf{adj1}}$**,*"ready to fall asleep"*. It would be safe to assume if someone is in a state of **sleep$_{\mathbf{n1}}$** then they are also engaged in **sleep$_{\mathbf{v1}}$**. However, it is arguably untrue that they are also **sleepy$_{\mathbf{adj1}}$**. Given this example, a reader should use derivationally-related links as evidence of an entailment relation rather than definite proof.
4. A reader learns a great deal more when considering several closely related definitions all at once instead of one at a time. Figure 2 shows some of the axioms we generate for concepts related to **waken$_{\mathbf{v1}}$**, *"cause to become awake or conscious"* and **waken$_{\mathbf{v2}}$**, *"stop sleeping"*. Notice that we do not find any entailments between **waken$_{\mathbf{v1}}$** and **waken$_{\mathbf{v2}}$** directly from the glosses. However, we have evidence that they are closely related considering that both are entailed by **reawaken$_{\mathbf{v1}}$** and both entail ¬**cause_to_sleep$_{\mathbf{v1}}$**. Given enough evidence we can conclude that **waken$_{\mathbf{v1}}$** and **waken$_{\mathbf{v2}}$** are different realizations of the same event.

Intuitions 1 through 3 are essentially ways of extracting more knowledge by exploiting conventions specific to concept definition. Intuition 4 provides an idea for combining the literal interpretation of a gloss with this extra knowledge to find even more concept knowledge. However, before we can do that we have to axiomatize the subtext we have extracted using intuition 1 through 3.

From intuition 1 we alter the semantics of relations of the form, e ENTAILS $\neg e'$ and add to it the following:

$\forall x.\ e'$ PRE $x \rightarrow e$ ENTAILS x

$\forall x.\ e'$ POST $x \rightarrow e$ ENTAILS $\neg x$

i.e. e occurring implies that e' is not occurring but its preconditions hold and its postconditions do not hold. In the case of **stay_up$_{\mathbf{v1}}$**, *"not go_to_bed$_{vl}$"* we add the axiom: $stay\_up_{v1}$ ENTAILS $\neg sleep_{n1}$

Intuitions 2 and 3 require weak axioms to represent evidence rather than certainty. To handle inference with these axioms we use Markov logic networks (MLNs) (Richardson and Domingos 2006), which have been shown to be an effective tool in natural language applications like semantic similarity and textual entailment (Garrette, Erk, and Mooney 2011; Beltagy et al. 2013). MLNs allow us to define both certain axioms (like those derived from interpreting glosses literally) which are analogous to first order logic, and weighted axioms that can be used to reason with uncertainty. Instead of an absolute truth value, we can infer probabilities for each predicate. We prefix weak axioms with $[w]$ to indicate they are true with some weight, w. From intuition 2, for every definite axiom we derive from the literal gloss interpretation we also add a weak converse. For $stay\_up_{v1}$ ENTAILS $\neg sleep_{n1}$ we create the axiom:

$[w]\ \forall t. \neg hold(sleep_{n1}), t) \rightarrow hold(stay\_up_{v1}, t)$

Likewise, for $waken_{v1}$ PRE $sleeping_{n1}$:

$[w]\ \forall t. \neg hold(sleeping_{n1}, t) \rightarrow stop(waken_{v1}), t)$

Intuition 3 simply adds a weak entailment axiom for each derivationally-related link, e.g. $[w]\ \forall t. hold(sleep_{n1}, t) \rightarrow hold(sleep_{v1}, t)$.

The weight of the soft axioms, w, indicates how closely we think glosses approximate their concepts in general. For our experiments we have found that $w = 2$ gives predictions that fit our intuitions. Changing w will alter the inferred probabilities; however, we will only use these probabilities to discriminate between actual commonsense knowledge and false positives. Instead of embodying real world probabilities, w should be thought of as a tuning variable.

Following from intuition 4, we combine several related axioms together, centered around a single concept, to form what we call a micro-theory. A micro-theory contains the axioms most salient to understanding its center concept. Given a micro-theory and a set of assertions about the the world, we can use the temporal model described in the previous section to infer the confidence the system has in a predicate being true at any time step. Figure 7 gives an example of the scores for a few predicates inferred using a micro-theory centered around **sleeping$_{\mathbf{n1}}$**.

The micro-theories may contain denser knowledge than current methods of knowledge extraction but they are unlikely to be used as commonsense KBs by the community. Outside users would have to commit to everything from ITL representation to the MLN inferencing when all they really want are simple relations, like $waken_{v1}$ ENTAILS $waken_{v2}$. If we want our KB to be useful then we need to distill all of these axioms into simplified relationships.

Inferring Semantic Relationships

We estimate the confidence our system has in the three relationships (ENTAILS, PRE, and POST) based on co-occurrence inferred from the axioms we extracted from definitions in the previous section. Figure 6 defines the confidence estimator, $\widehat{C}$, for each relationship. If we wanted to find the value of $sleeping_{n1}$ ENTAILS $snore_{n2}$ we would first build a micro-theory centered around **sleeping$_{\mathbf{n1}}$**. Next we create a set of premise predicates (which we will call observations) about **sleeping$_{\mathbf{n1}}$**, like $\{hold(sleeping_{n1}, 1), \neg hold(sleeping_{n1}, 2)\}$, then run inference using the micro-theory and the observation set to obtain the probability scores in Figure 7. We average the confidence that $snore_{n2}$ holds at time t, given that $hold(sleeping_{n1}, t)$ was in the set of observations. In this case we have only one instance, $P\big(hold(snore_{n1}, 1)\big) = .5$ and so, $\widehat{C}(sleeping_{n1}$ ENTAILS $snore_{n2}) = .5$. Likewise, $\widehat{C}(\neg sleeping_{n1}$ ENTAILS $snore_{n2}) = 0$.

In general, for $\widehat{C}(xRy)$, a score of 1.0 means the systems is very confident in xRy, 0.5 means it has no knowledge of xRy, and 0 means it is very confident in $xR\neg y$. We interpret these numbers to mean that knowing **sleeping$_{\mathbf{n1}}$** holds

1. $\widehat{C}(s \text{ ENTAILS } s') =$
   $\langle\{P\big(hold(s',t)\big) \mid hold(s,t) \in OBS\}\rangle$
2. $\widehat{C}(e \text{ ENTAILS } e') =$
   $\langle\{P\big(occur(e',t)\big) \mid occur(e,t) \in OBS\}\rangle$
3. $\widehat{C}(e \text{ PRE } s) =$
   $\langle\{P\big(hold(s,t)\big) \mid start(e,t) \in OBS\}\rangle$
4. $\widehat{C}(e \text{ POST } s) =$
   $\langle\{P\big(hold(s,t)\big) \mid stop(e,t) \in OBS\}\rangle$

Figure 6: Confidence estimators ($\widehat{C}$) for each temporal relation. Each relation $x\ R\ y$ is estimated by the average confidence score that y is in a certain state at the same time we have asserted something about x.

does not indicate the state of **snore$_{n2}$**. But, we are very confident that $\neg sleeping_{n1}$ ENTAILS $\neg\, snore_{n2}$. We can generate a KB by adding all relations with confidence scores that pass some threshold (e.g. $\widehat{C} < .15$ or $\widehat{C} > .85$)

We want to know if a relationship is true in general, regardless of starting conditions. But these numbers are directly influenced by the observation set that generated them. Different initial assertions could predict very different things. Consider $\{hold(sleeping_{n1}, 1), hold(sleeping_{n1}, 2)\}$ and $\{hold(sleeping_{n1}, 1), \neg hold(sleeping_{n1}, 2)\}$. In the first set, **sleeping$_{n1}$** is static so we would only learn about what happens when **sleeping$_{n1}$** is true. But in the second set **sleeping$_{n1}$** changes which entails some event has occurred which may have complex consequences.

To get a better idea of **sleeping$_{n1}$**'s relationships we have to look at the confidence scores given from several different starting assertions. For each concept, the system generates all combinations of it holding or occurring (depending on its type) in scenarios with up to three time steps (see Figure 8). For instance, from **sleep$_{n1}$** we would generate two scenarios each with a single time step, four scenarios with two time steps each, and eight with with three time steps: $\{hold(sleep_{n1}, 1)\}$, $\{\neg hold(sleep_{n1}, 1)\}$, $\{hold(sleep_{n1}, 1), hold(sleep_{n1}, 2)\}$, $\{hold(sleep_{n1}, 1), \neg hold(sleep_{n1}, 2)\}$, etc. We use the inference results from these scenarios to calculate the confidence our system has in relations of the form $sleep_{n1}\, R\, x$.

Evaluation

To test our framework, we generated micro-theories for 170 concepts related to **sleep$_{n1}$**, **die$_{v1}$**, **cure$_{v1}$**, **clean$_{adj1}$**, **wet$_{v1}$**, **swimming$_{n1}$**, and **ignition$_{n3}$** (in the sense of burning). Each

$P(hold(snore_{n1}, 1)) = .5$ $\quad P(hold(snore_{n1}, 2)) = 0$
$P(occur(waken_{v1}, 1)) = .8$ $\quad P(stop(waken_{v1}, 2)) = .9$

Figure 7: Inferences for some predicates made using a micro-theory about **sleeping$_{n1}$** with the assertions: $hold(sleeping_{n1}, 1)$ and $\neg hold(sleeping_{n1}, 2)$.

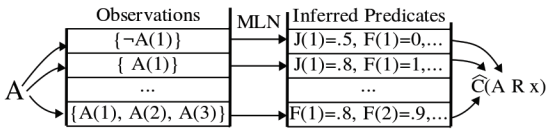

Figure 8: The process of estimating relationship confidence. The system builds several observation sets and calculates $\widehat{C}(A\, R\, x)$ from the probabilities inferred from those sets.

micro-theory involves anywhere from 70 to 450 WN concepts.

To evaluate the four types of relationships shown in Figure 6, we sampled 76 relationships from each type (304 in total), a third of which had confidence scores between .2 and .8 and the other two-thirds of which were drawn from outside that range. We did this because we are more interested in testing the relations the system was confident about.

From each of the relations we sampled, we created a multiple choice question. Each question presents a sentence describing a relationship between two concepts and asks the respondent to fill in a blank with either, "definitely", "maybe (unsure)", or "definitely not" For instance, to test an ENTAILS relation between a sense of **sleeping** and a sense of **asleep**, the question would read: "If something is sleeping then it is _____ asleep." The glosses for each of the concepts are also provided but the respondent is instructed to answer using their own knowledge. We found that crowdsourced respondents tended to misunderstand entailment relationships so we answered the questions ourselves.

We evaluate our results by first converting system confidence scores to responses. If $\widehat{C}(xRy) \geq .85$ then the system answers, "definitely", if $\widehat{C}(xRy) \leq .15$ then it answers "definitely not", otherwise the system answers "maybe". The system's response is correct if it matches the human's answer. Since precision is so important to the KB building task, we are only interested in relations the system is very confident about. To that end, we do not evaluate any questions the system answered with "maybe". After filtering out those system responses, we are left with about 50 questions. Figure 9 shows how changing the threshold for answering affects the system's performance. From the 170 concepts we looked at, we found over 3400 relations with a probability that pass the .15–.85 threshold.

The precision, recall, and $F_{.5}$ scores are presented in Table 1. To calculate recall, we counted all questions where the correct response passed the .15–.85 threshold and divided it by the number of human responses that were "definitely" or "definitely not". There are about 160 such human responses. We calculate the $F_{.5}$ score to weight precision higher than recall because when building a lexical KB, we are more concerned with the quality of knowledge than the coverage. To judge the effects of subtext knowledge, we also present results using a model without any uncertain axioms (No Soft Axioms).

As shown by the results, the full model's precision score is lower than the ablated model's score; however, the recall

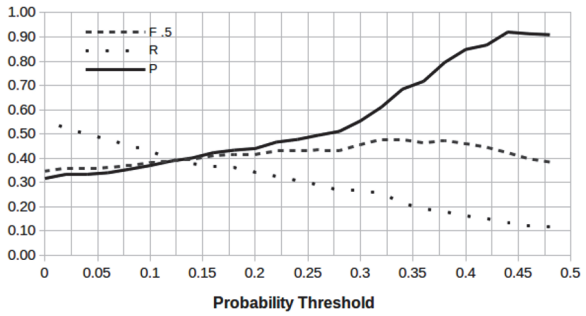

Figure 9: A plot showing how precision, recall, and $F_{.5}$ are affected by the probability threshold. Everything that falls outside .5 $\pm$ *threshold* is evaluated.

Full			No Soft Axioms		
P	R	$F_{.5}$	P	R	$F_{.5}$
0.71	0.20	0.47	0.91	0.12	0.38

Table 1: Stats for two models. **Full** uses all the features described in the paper. **No Soft Axioms** does not use uncertain converses of gloss axioms or derivationally related axioms.

and the $F_{.5}$ scores are higher. In terms of raw numbers, the full system found 3400 concept relations while the ablated model only generated 1000 relations. Even when factoring in the precision scores, the full model generates substantially more valid concept relations.

Omitting the uncertain axioms (e.g. $[w]LF_{Gloss} \rightarrow Concept$) sacrifices recall for precision. Soft converse axioms provide more opportunities for inference and therefore a higher recall; however, these inferences are based on assumptions which could be too strong or outright wrong. Although precision is ultimately the most important score when building a KB, it defeats the purpose if the knowledge is too sparse. In the future we will tune w to home in on the ideal weight to apply to our assumptions.

Conclusion

We have identified heuristics along with an inference process that can find new relations among WN concepts that existing methods ignore. By axiomatizing concept definitions and their subtexts and probing their predictions about co-occurrence, we were able to infer temporal and causal relationships that were not explicitly represented in WN. The results indicate that this framework could augment lexical KBs with knowledge that is difficult to find by other means. We are currently exploring features that predict overly confident inferences so we can further raise precision without sacrificing recall.

Acknowledgments

This work was supported in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO).

References

Allen, J. F., and Ferguson, G. 1994. Actions and Eventis in Interval Temporal Logic. *Journal of Logic and Computation* 1–56.

Allen, J.; de Beaumont, W.; Galescu, L.; Orfan, J.; Swift, M.; and Teng, C. M. 2013. Automatically Deriving Event Ontologies for a Commonsense Knowledge Base. In *Proceedings of the International Conference for Computational Semantics*.

Allen, J. F. 1984. Towards a general theory of action and time. *Artificial Intelligence* 23(2):123–154.

Allen, J. 2014. Learning a Lexicon for Broad-coverage Semantic Parsing. *Proceedings of the ACL 2014 Workshop on Semantic Parsing* 1–6.

Beltagy, I.; Chau, C.; Boleda, G.; Garrette, D.; Erk, K.; and Mooney, R. 2013. Montague Meets Markov: Deep Semantics with Probabilistic Logical Form. *Joint Conference on Lexical and Computational Semantics (*SEM)* 11–21.

Carlson, A.; Betteridge, J.; Kisiel, B.; and Settles, B. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*.

Clark, P.; Fellbaum, C.; Hobbs, J. R.; Harrison, P.; Murray, W. R.; and Thompson, J. 2008. Augmenting WordNet for deep understanding of text. *Proceedings of the 2008 Conference on Semantics in Text Processing - STEP '08* 45–57.

Garrette, D.; Erk, K.; and Mooney, R. J. 2011. Integrating Logical Representations with Probabilistic Information using Markov Logic. *In Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)* 105–114.

Grice, H. P. 1975. Logic and conversation. Speech acts, ed. by Peter Cole and Jerry Morgan, 41-58.

Harabagiu, S.; Miller, G.; and Moldovan, D. 1999. WordNet 2 - A Morphologically and Semantically Enhanced Resource. *Proceedings of SIGLEX-99* 1–8.

Kim, G., and Schubert, L. 2016. High-Fidelity Lexical Axiom Construction from Verb Glosses. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* 34–44.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. J. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4):235–244.

Orfan, J., and Allen, J. 2015. Learning New Relations from Concept Ontologies Derived from Definitions. *2015 AAAI Spring Symposium Series* 1–4.

Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine learning* 62(1-2):107–136.

Stock, P. 1988. The structure and function of definitions. In *ZuriLEX'86 proceedings: Papers read at the EURALEX International Congress, University of Zürich 9-14 September 1986*, 81–89.