WordNet for "Easy" Textual Inferences

Aikaterini-Lida Kalouli, Livy Real and Valeria de Paiva
University of Konstanz, University of São Paulo, Nuance Communications
aikaterini-lida.kalouli@uni-konstanz.de, livyreal@gmail.com, valeria.depaiva@gmail.com

Abstract

This paper presents a WordNet-based automatic approach for calculating "easy" inferences. We build a rule-based system which extracts the pairs of the SICK corpus whose sentences only differ by zero or one word and then identifies which inference relation (i.e. entailment, contradiction, neutrality) exists between these words, based on WordNet relations. Since the sentences of those pairs only differ by the words of the comparison, the inference relation found between the words is taken to apply to the whole sentences of the pair. For some cases not dealt by WordNet we use our own heuristics to label the inference type. With this approach we accomplish three goals: a) we manage to correct the annotations of a part of the SICK corpus and provide the corrected corpus, b) we evaluate the coverage and relation-completeness of WordNet and provide taxonomies of its strengths and weaknesses and c) we observe that "easy" inferences are a suitable evaluation technique for lexical resources and suggest that more such methods are used in the task. The outcome of our work can help improve the SICK corpus and the WordNet resource and it also introduces a new way of dealing with lexical resources evaluation tasks.

Keywords: WordNet, natural language inference, SICK corpus, evaluation of lexical resources

1. Introduction

"Understanding entailment and contradiction is fundamental to understanding natural language, and inference about entailment and contradiction is a valuable testing ground for the development of semantic representations" say Bowman, Angeli, Potts and Manning in their introduction of SNLI (Bowman et al., 2015), the Stanford Natural Language Inference corpus. We agree and also share their goal of providing semantic representations for sentences which can then be used to compute inference relations between them. To reach this goal we started by investigating SICK by Marelli et al. (2014b), an inference geared corpus that we would like to use as the golden standard for our inference system. This investigation led us to interesting observations on the logic of contradictions, shed light onto faulty corpus annotations and gave us insights for the task at hand, as we discuss in Kalouli et al. (2017b) and Kalouli et al. (2017a). In this previous work we attempted to correct some of those faulty annotations but we soon realized that we could not manually check and correct the whole SICK corpus in a reasonable amount of time. Therefore, we decided to find better ways of correcting sub-parts of the corpus, which led us to this work.

In this work we present our automatic approach for re-annotating and thus correcting a subset of the SICK corpus. The approach is strongly based on Princeton Wordnet (PWN) (Fellbaum, 1998). But the corrected sub-corpus we get as outcome of this work is only one of the three contributions of this paper. Additionally, the approach can be used as a good preliminary basis for identifying "easy" inferences, meaning inferences where syntax is ruled out as a "common denominator" and a sentential inference boils down to a lexical inference and to the one-to-one lexical semantic mappings of the words involved. In other words, what we call "easy" inferences here, are pairs of sentences that can be labelled for entailment/neutral/contradiction relations considering only lexical semantics or world-knowledge. Identifying which inferences are "easy" and how many of them can be achieved with existing lexical tools is important if we want to pursue our goal of computing complex inferences. We believe that complex inferences can be broken down to easy ones and that we need to know how to handle the easy ones first. We also believe that for a symbolic grounded inference system it is important to distinguish different phenomena that play a role in Natural Language Inference (NLI) tasks and then have different ways to deal with them, as pointed out by McCartney (2009). With this approach, we thus seek to evaluate the completeness of PWN as a lexical resource for inference and identify strengths and weaknesses of the lexicon which can be used to improve the resource. A successful evaluation will bring us to our last goal which is to propose that such "easy" inferences tasks are good evaluation methods for lexical resources and that such methods should be used more often as one type of evaluation of appropriate lexical resources. Evaluating lexical resources from a qualitative point of view, more than simply in terms of coverage numbers, is a well known and still open issue for the Lexical Resources community, as pointed out, e.g., by de Paiva et al. (2016).

In the following section we will briefly introduce the SICK corpus. In section 3. we will describe in detail the approach we developed and how it helps us to automatically correct a part of the corpus. In the section after we will evaluate our approach by offering the results of our manual investigation and providing a taxonomy of "easy" inferences found in SICK. In section 5. we will discuss in detail the threefold contribution of this approach and how it can be used further. In the last section we will offer some conclusions and plans for future work.

2. The SICK corpus

SICK (Sentences Involving Compositional Knowledge) by Marelli et al. (2014b) is an English corpus, created to provide a benchmark for compositional extensions of Distributional Semantic Models (DSMs). The data set consists of

English sentence pairs, generated from existing sets of captions of pictures. The authors of SICK selected a subset of the caption sources and applied a 3-step generation process to obtain their pairs. This data was then sent to Amazon Turkers who annotated them for semantic similarity and for inference relations, i.e. for entailment, contradiction and neutral stances. Since SICK was created from captions of pictures, it contains literal, non-abstract, common-sense concepts and is thus considered a simple corpus for inference. The corpus is *simplified* in aspects of language processing not fundamentally related to composionality: there are no named entities, the tenses have been simplified to the progressive only, there are few modifiers, etc. The curators of the corpus also made an effort to reduce the amount of encyclopedic world-knowledge needed to interpret the sentences.

The data set consists of 9840 sentence pairs, which have been annotated as 1424 pairs of contradictions ($AcBBcA$), 1300 pairs of double entailment ($AeBBeA$), 1513 pairs of entailment ($AeBBnA$) and 4992 pairs of neutrals ($AnBBnA$). The SICK corpus is a good dataset to test approaches to semantic representations and natural language inference, due to its intended, human-curated simplicity; the pairs talk about everyday, concrete actions and actors. The fact that the captions were produced by different humans, should provide us with near paraphrases or different ways of describing the same scene. The process of normalization added some of the inferences that the corpus was meant to capture, e.g. negations and modifier dropping inferences were added. The number of sentences pairs of the corpus may seem substantial (almost 10K of pairs), but there is much redundancy in the corpus. In total we have 6076 unique sentences and only around 2000 unique lemmas, which means a few more concepts, as assigned by PWN synsets.

3. The “one-word difference” approach

Our approach of automatically annotating and correcting the inference pairs is based on the observation that several SICK pairs only differ by none or one word. Differing by “one word” means that there is either one more word in the one sentence than in the other or that each of the sentences contains a word that is not found in the other one. (We say two sentences differ by “no word” when they differ only in their use of the determiners *the* and *a/an*.) They are thus the perfect ground for dealing with some “easy” inferences, as we would like to call them, because we can ignore the syntax involved and find the relation between the pairs solely based on the relation between the different words. A nice example of a “one-word difference” pair is *Kids in red shirts are playing in the leaves.* vs. *Children in red shirts are playing in the leaves*, where the only difference is *kids/children*. This approach can automatically correct and re-annotate some of the pairs without having to solve all the inference challenges associated with the meanings of the sentences first. The approach will become clearer in the following.

3.1. Processing SICK

We parsed the SICK corpus sentences with the Stanford Enhanced Dependencies (Schuster and Manning, 2016), which offer us a strong basis for further processing. Then, the sentences were run through the knowledge-based JIGSAW algorithm (Basile et al., 2007) which disambiguates each (noun, verb, adjective, adverb) word of the sentence by assigning it the sense with the highest probability. Briefly, JIGSAW exploits the WordNet senses and uses a different disambiguation strategy for each part of speech, taking into account the context of each word. It scores each WordNet sense of the word based on its probability to be correct in that context. The sense with the highest score is assigned to the word as the disambiguated sense. Using this PWN-based algorithm corresponds to using PWN as our basic ontology or knowledge graph for the approach implemented. Princeton WordNet is a basic ontology and we expect that many inferences will not be supported. However, it is surprising how much we can get from it, which shows that, for the task at hand, PWN has the coverage we need (a similar sort of phenomenon, where PWN worked better than a more traditional ontology, Cyc, was observed in de Paiva et al. (2007). However, on that setting, much more information was available from the syntax, which was based on the Xerox Language Engine (XLE) and Lexical Functional Grammar (LFG) f-structures.)

3.2. Finding the “one-word difference” pairs

Having done this shallow linguistic processing of the sentences, we now focus on the surface form of the sentences and extract the ones that differ by none or only one token (we will call these “words-apart” from now on). Since we started working on the surface level, one should note that e.g. *drum* and *drums* still count as different words at this point. We create a small module which takes as input each pair of SICK and checks if the sentences of the pair contain more than two different words. This works on the basis of the creation of sets of words out of the two sentences and the comparison of the sets. If the sets have more than two different words, then they are discarded; if they are different by none, one or two (one from each sentence) words, then the pair is written in a new file, along with the words by which the pair is different as well as which sentence each of the “words-apart” comes from (e.g. the pair *A= A person in a black jacket is doing tricks on a motorbike. B= A man in a black jacket is doing tricks on a motorbike* would be assigned the pair of words *A:person,B:man*).
Note that we choose to exclude some determiners from this comparison. As discussed in Kalouli et al. (2017a), we need to take the SICK pairs as referring to the same entities and events, no matter if the introducing determiners are definite or indefinite articles, to be able to compute contradictions. Since we assume co-reference no matter the definiteness of the articles in the sentences of the pair, we can also exclude them from the difference comparison so that they do not count as words by which the sentences could be different. Note that this approach does not exclude all determiners from the corpus, but only the determiners *the* and *a*. Other determiners that play a role in SICK relations, as well as quantifiers, are taken into account.

By running this module on all 9840 pairs of SICK, we end up with 2936 pairs being "one-word apart" [1], so almost 30% of the corpus.

3.3. Assigning relations to the pairs

The two previous processing steps are necessary for the step of automatically assigning inference relations to the "one-word difference" pairs. We create a second module that takes the "words-apart" of the previously extracted pairs and depending on the nature of those words it either runs some heuristics on them or feeds them to WordNet for further processing.

Heuristics for non-lexical relations If at least one of the "words-apart" is not a PWN word, i.e. a noun, a verb, an adjective or an adverb — in other words, if it is one of the word classes not handled by PWN — then the "words-apart" are fed into a heuristic engine that decides which label should be given to the pair.

We need such an engine to account at a very primitive level for the missing syntax and at the same time to not lose the precision of such pairs. Only the following cases are dealt with:

- one of the words is a form of the auxiliary *be* (the only one used in SICK) and the other one is the negated version of that auxiliary: the sentences contradict each other;
- one of the words is the negation particle *not* or *no*: the sentences contradict each other;
- there is only one different word and it is the quantifier *one* which is handled as a determiner and thus "ignored" (see section 3.2.): the sentences entail each other;
- the two words are opposing prepositions, e.g. *on-off*, *up-down*, *with-without*, *in-out*: the sentences contradict each other[2];
- both words are quantifiers or there is only one different word and it is a quantifier: depending on the quantifiers different heuristics apply; e.g. if the word of A is the quantifier *many* and the word of B the quantifier *few*, then the sentences contradict each other but if the word of A is the quantifier *many* and the word of B the quantifier *some*, then sentence A entails sentence B but sentence B is neutral to A, etc;
- both words are one of the pronouns *someone, somebody* or one word is one of those and the other one is the word *person*. In both cases the sentences are taken to entail each other.[3]

This means that every pair that enters this engine is finally labelled with one of the inference relations $AeBBeA$, $AeBBnA$, $AcBBcA$, $AnBBnA$, $AnBBeA$ or "-", where A stands for sentence A, B for sentence B, e for $entails$, c for $contradicts$ and n for $neutral$. We use the symbol "-" for cases the heuristics cannot deal with.

WordNet for lexical relations If none of the "words-apart" is one of the above cases, then the words are fed into our PWN-based mechanism. The mechanism retrieves from our local repository of PWN3.0 the synonyms, hypernyms, hyponyms and antonyms that correspond to the disambiguated sense of each word, as this was assigned during the step of processing SICK with JIGSAW (see Section 3.1.). The entries found for each lexical relation (i.e. synonymy, hypernymy, hyponymy, antonymy) of the one word are compared with the entries for each lexical relation of the other word. Depending on the ordering of the sentences within the pair, different monotonicity rules apply (Hoeksema, 1986). For example, if the word A is one of the hyponyms of the word B, then there is upward monotonicity that implies that sentence A will entail sentence B but B will be neutral to A. Similarly, if the word A is one of the synonyms of the word B, then the two sentences entail each other. The mechanism takes into account all possible combinations between the lexical relations of the "words-apart" and gives to each pair one of the inference labels mentioned above. If no relation between the "words-apart" can be established, then the pair is left unlabelled. If one of the "words-apart" cannot be found within PWN altogether, then the pair is marked with the label "not found".

The senses contained in SICK are expected to be daily actions and common entities that a knowledge base like PWN should already have. (By contrast, in a more specialized corpus such as a biomedical one, we would expect to need to add to the standard English vocabulary, the specific biomedical vocabulary required by the application.) We expect, for example, that the lexical resource knows that a *dog* is an *animal*, an easy and obvious taxonomic inference. After comparing some of the words found in SICK as a whole with the ones contained in PWN3.0, we observe that some words or senses are still missing. For instance, PWN has no adjective *shirtless* nor the noun *footbag*, although they are established dictionary words. Concretely, we observed that some 15 nouns are missing from PWN3.0. For the 1100 unique nouns of SICK, lacking only so few shows that PWN has a large coverage of English concepts and can be used for a corpus like SICK. However, we must remember that SICK is simplified on purpose, it aims to not have multiword expressions (MWEs), named entities or compounds. This is an important characteristic of the corpus that provides us with good results in this task. It is well-known that WordNet misses many of the well established MWEs in English, which may mean that, if we want to deal with larger inference corpora, like SNLI, we should extend our

[1] Available under `https://github.com/kkalouli/SICK-processing/tree/master/word_difference/one_word_difference`

[2] Princeton WordNet contains no functional words, e.g. no prepositions nor pronouns, so it cannot deal with meanings that depend on them. Newer work from the ARK Lab in CMU provides meanings for prepositions, so we hope to investigate the use of their resources described in `http://www.cs.cmu.edu/~ark/` soon and perhaps integrate some more of them in this module.

[3] We use this heuristic because, since PWN has no pronouns, *someone, somebody* are not mapped to the concept of *person* as humans would naturally map them.

resources using perhaps Wiktionary and Wikipedia. Even for SICK processing, WordNet lacks some concepts; it does not have *jetski* nor *jet ski* or even *water ski*, for instance. It does not have nouns such as *motocross, wetsuit, corndog* or verb predicates like *rock climb, unstitch, wakeboard* (verb). Other concepts of SICK cannot be found in PWN because of tokenization issues. Wordnet lists *fistfight* instead of *fist fight*, and *ping-pong* instead of *ping pong*, for instance, but SICK uses the tab-separated notation so the concepts do not match. Although it might sound trivial, this inconsistency causes several mappings to fail. Additionally, despite trying to avoid compounds, SICK has 1129 of these, as counted based on the Stanford dependencies. These come down to 435 unique compounds. Out of the 435 unique compound nouns in the processing of SICK, only 84 are included in PWN. Of course, many might not deserve to be listed as compounds in PWN. The criteria to be used for dictionarizing a compound is a thorny subject. For instance, a *toy train* is a perfectly compositional compound that appears in Wikipedia. Lexicographers perhaps have no need to list these compositional compounds, but ontologists (especially the ones interested in massive processing of texts) need to do so.

Our PWN-based mechanism has the merit of precision. No matter if ten or a hundred Turkers say that a *man* and a *person* entail each other, PWN will tell us that men are persons, but there are other persons too. So the sentence *A man in a black jacket is doing tricks on a motorbike* entails the sentence *A person in a black jacket is doing tricks on a motorbike*, but not conversely. Similarly, PWN will also tell us that a *guitar* is a musical *instrument*, but not all instruments are guitars and thus it avoids the issue noted by Beltagy and described in `https://github.com/ibeltagy/rrr` [4], that makes guitars and flutes entail each other. Note that this theoretical precision can be broken if the tools on which our system is based, i.e. the Stanford Parser and the JIGSAW algorithm, deliver faulty output. For instance, a missrecognized part-of-speech will lead to a faulty disambiguation which might lead to the assignment of the wrong PWN label.

But in this paper we wish to examine concretely what is the coverage of WordNet for the "one-word difference" pairs and not for the whole SICK corpus. In the next section we will evaluate our approach and discover strengths and weaknesses of this approach and WordNet.

4. Evaluation of the approach

The "one-word difference" approach presented above was applied on all 2936 pairs that are "one-word" apart and it could automatically label 1651 of them. We manually looked at both the labelled and unlabelled pairs to see on the one hand if the labelled pairs have the right annotation and on the other hand which kinds of lexical inference can be accomplished by PWN and which senses or relations are still missing.

4.1. Evaluation of WordNet labelled pairs

Our manual investigation of all 1651 pairs showed us that our "one-word difference" approach is reliable and has an almost 100% accuracy as it will be shown shortly. Although not all pairs get a label, the 1651 that do, are assigned mostly the correct inference relation.

We could confirm 1100 contradictions with most of them coming from the non-lexical heuristics we defined and 200 coming from lexical antonyms. We additionally found 179 single-sided entailments which correspond to hypernymy and hyponymy relations, two of the main PWN relations. These are taxonomic subsumptions of the kind: a *dog* is an *animal*, the collection of *pianists* is contained in the collection of *persons* and a *man* is a *person*.

We also have 330 double entailments coming mostly from synonyms known to PWN, e.g *couch* and *sofa*, *clean* and *cleanse* or *carefully* and *cautiously* or from some of our heuristics, e.g. the quantifiers heuristics. There are 199 pairs out of these double entailments which belong to a third category, in which no different word is found within the pair, e.g. *A = The teenage girl is wearing beads that are red. B= A teenage girl is wearing beads that are red.* However, since the very basic processing we are doing only considers the surface forms of the sentences and cannot distinguish between agents and patients, 33 pairs out of the 199 are wrong because the order of the words is changed, causing the predicate arguments to be scrambled and thus the sentences to not entail each other, e.g. *A= A baby is licking a dog. B= A dog is licking a baby.* These 33 pairs (1,9%) out of the 1651 labels cost us the 100% accuracy.

Using the present approach, we could automatically correct pairs such as *A = A woman is combing her hair. B= A woman is arranging her hair* that was labelled as $AnBBnA$ in the original SICK and in our present version in annotated as $AeBBnA$. In this way, we can improve the human annotation.

4.2. Evaluation of unlabelled pairs

There were 1285 pairs that could not get a PWN label (cf. Table 1). Surprisingly, only a few of them were due to words missing altogether from PWN; the rest were due to missing relations between the terms. The words *debone, atv (all terrain vehicle), biker and kickboxing* for example are missing from PWN3.0 altogether. A few other failures are due to issues with the disambiguation. For example, for the pair *A = A woman is amalgamating eggs. B= A woman is mixing eggs*, PWN does have the verb *amalgamate* in the same synset as *mix*, but JIGSAW wrongly assigns *amalgamate* to the lemma *amalgam* and wrongly annotates it as an adjective and thus as such cannot find it within PWN.

We have 325 pairs that we annotated as antonyms or near antonyms. Knowing that the corpus was constructed aiming for a reasonable number of contradictions and assuming that sentences refer to the same events and entities, we believe pairs such as *Children in red shirts are playing in the leaves* and *Children in red shirts are sleeping in the leaves* need to be annotated as contradictions, although *sleep* and

[4] [...] This is because of inconsistencies in the annotations of the SICK dataset (remember that most of the rules are automatically annotated using the gold standard annotation for the pair where the rule is extracted from). For example, the relation between "flute" and "guitar" could be Entail but in most cases it is Neutral.

play are not direct antonyms. The same children cannot be sleeping and playing at the same time. These intended contradictions account for a high number of the illogical annotations we have observed before in Kalouli et al. (2017b). This pair was annotated by Turkers as $AnBBcA$, instead of $AcBBcA$. But such an antonym relation is not present in PWN. It is world knowledge that people, even kids, cannot play and sleep, or sit and jump at the same time. Many of the 325 pairs can be accounted for by such world knowledge. It is an interesting, open question whether some of these relations should be included in PWN and if yes, under which category. Some other *near antonyms* bring us to the well-known difficult issues of deciding on the granularity of events: *A man is resting* is not contradictory with *A man is exercising*, but the same man at the same moment cannot be doing both, even if exercising requires some resting between exercises.

There are 299 pairs that we called 'intersective'. These correspond to a single word difference and this word, usually either an adjective or an adverb, provides an intersective subset of the predicate described. For example, in the pair *A skilled person is riding a bicycle on one wheel. A person is riding a bicycle on one wheel*, we only need to check that a *skilled person* is a *person*. Similarly for the example *Some fish are swimming quickly. Some fish are swimming* we only need to know that *swimming quickly* implies *swimming*. A few of these intersectives are actually compounds, like *swimming pool, cyclone fence, etc*. Such 'intersective' cases are not expected to be handled by PWN as they need a module for inference, even if just a basic one, to deal with them. This example confirms what we pointed out in the introduction: even such "easy" inferences pose challenges and are not as "easy" as one might expect and therefore we need to be able to do these first, if we really want to compute more complex inferences.

Moving on with our investigation, among the unlabelled pairs, we found 283 that belong mostly to the taxonomic relations we described before, i.e. hyponymy/hypernymy and synonymy, and would thus be single-side (259 pairs) or double entailments (24), respectively. On the one hand, this (positively) low number (24) of double-entailments, or synonyms, not labelled by PWN shows interesting weaknesses of PWN. For example, PWN has nine synsets for the verb *fire*, at least four of which (02002410, 01133825, 01135783 and 01134238) have to do with guns and weapons, but the verb *shoot* does not appear anywhere in these four synsets. Similarly, the noun *cord* has four synsets, only one (04108268) relevant to its similarity to *rope*, which also has four noun synsets, only one relevant to *cord* (03106110), but these two synsets are not connected at all. On the other hand, the higher number of single-side entailments left unlabelled can mainly be explained by more complex challenges than plain weaknesses of PWN. For example, *to perform* does not necessarily imply *to play*; one can perform mimes, act on plays, do performance art. But *A band is performing on a stage* does entail that *A band is playing on a stage* and conversely. So, again here, we have relations, that only work in the specific context of the other arguments provided, similarly to what we observed for the antonyms.

It is again worth discussing if and how such relations and information should be encoded in lexical resources such as PWN. For some of them, we are convinced that we will need to use the strengths of machine-learning and word embeddings, which could probably give us some of the intended relations; e.g. in the pair *The dog is catching a black frisbee. The dog is biting a black frisbee*, the words *catch* and *bite* describe pretty different actions but in the context of a dog, the words are to be treated as similar. We have also observed that such harder cases mostly involve verbs as their senses are more controversial than nouns.

The further categories discussed in what follows constitute smaller groups. Firstly, there are 27 pairs whose sentences involve meronymy relations and precisely what specific nouns are made of. A representative example is the sentence *A dog is running on the beach and chasing a ball* pairing to *A dog is running on the sand and chasing a ball*. Since our approach is not considering the meronymy relation of PWN, which would provide us with the information that a beach is made of sand, such cases remain unlabelled. Secondly, there is a collection of pairs (112) that seem to us a misguided effort on the part of the corpus creators to paraphrase certain complex expressions.

The first case (27 pairs) is the one of removing adjective expressions from the sentences. Transforming the sentence *A man in a black jersey is standing in a gym* into *A man in a jersey which is black is standing in a gym* seems a confusing source of mistakes for annotators and parsers.

The second case (32 pairs) is doing a similar job of rewriting 'noun-noun' compounds, but without creating a relative clause. For example, the sentence *A soccer player is scoring a goal* was expanded to *A player of soccer is scoring a goal* but how often would we say *player of soccer* instead of *soccer player*? These pairs mostly use the prepositions *for, of, from*, as in *fishing rod, roof top, tap water*, respectively: *a rod for fishing, the top of the roof, water from a tap*. Lastly, there are several pairs (53) where the expansion tried to explain a compound, to provide a definition for the term. To make it clearer, we can look at an example. The sentence *The crowd is watching two racing cars that are leaving the starting line* was paired to *The crowd is watching two cars designed for racing that are leaving the starting line*, in which there is an attempt to explain *racing cars* as *cars designed for racing*. But many other, less complex, closer to real-world 'definitions' could have been provided instead.

Clearly some of this information is lexical and could be codified having more Wikipedia-style world knowledge in PWN, like saying *motocross bike* is a kind of motorcycle for racing on dirty roads or a *ceiling fan* is a fan usually attached to the ceiling. Other information is instead the kind of world knowledge that tends to be codified in a knowledge base such as SUMO (Niles and Pease, 2001), like the fact that a *a sewing machine* is a machine used for sewing fabric and could thus not have been labelled by PWN anyway. However, many of the pairs of this category explain colors of concrete nouns, such as *blue shirt, brown duck, black dog* described as *a shirt dyed blue, a duck with brown feathers, a dog with a black coat*, respectively, which should be neither in the lexicon nor in a knowledge base in any case and it is thus not surprising that they are not found by PWN.

Near Antonyms	325
Intersective	299
Synonyms	24
Hypernyms/Hyponyms	259
Meronyms	27
Paraphrases	112
Dropping	34
Scramble	55
Similar	36
Others	114
Total	**1285**

Table 1: Phenomena in non-labelled pairs

Thirdly, we have 34 pairs of dropping modifiers or dropping conjunctions, for instance *A man is playing a piano at a concert. A man is playing a piano* or *The man is singing and playing the guitar. A man is playing a guitar*. Although such pairs can be solved by simple logic, similar to the one presented for the 'intersective' pairs, the knowledge required to do so is not lexical and is thus not encoded in PWN. Again, here we would need a basic inference module to do such "easy" inferences.

Additionally, we have 'scrambled' pairs (55), as described in Section 4.1.. Pairs such as *The woman is drawing a man. A man is drawing* cannot be resolved by lexical knowledge alone but instead would need at least a notion of comparing relationships.

Furthermore, we have cases of lexical similarity that are not really logical. For example, consider the pair *A= A dog is licking a toddler. B= A dog is licking a baby*. Toddlers are not babies, the words are not synonyms, but they are similar enough that people will use them as if they were synonyms. These similarity cases are interesting, as they prompt the question of how this kind of information should be encoded, similar to the discussion about "context-dependent" antonyms and synonyms early on. State-of-the-art machine learning techniques might be able to give us more expanded or more context-specific semantics for certain words which might facilitate this task.

Last but not least, there are cases of unlabelled prepositions, quantifiers and inter alia. As explained earlier, we only have a few prepositions in our heuristics and thus there are 42 pairs, whose differences are prepositional but our approach does not handle. Expansion of the heuristics would decrease this number. There are some 20 pairs that differ by numbers or quantifiers (e.g. *Three women are dancing. A few women are dancing*), for which more than lexical knowledge is required and another 40 pairs that seem to us really neutral and no linguistic knowledge, lexical or otherwise, would help. Representative is the pair *A man is thinking. A man is dancing*. People can dance and think at the same time. We call this entire last category 'Others' in the table following.

Looking at Table 1 it is clear that lexical semantics can only help with some of the phenomena, as it was described in detail above.

5. Contributions of the approach

As it was mentioned in the introduction, with the work and approach in this paper we hope to achieve three goals. In the following, we will see how each of these goals is fulfilled. We should note that this approach is simple, yet wide enough to be used on other corpora than SICK and achieve similar goals. Any corpus containing pairs of sentences differing by two or less words can be used as an application platform of this approach. There is nothing SICK-specific in this approach which makes the method ideal for verifying the annotations of further corpora, further evaluating WordNet and further discovering "easy" inferences. We see that similar efforts like the one by Pavlick and Callison-Burch (2016) are also breaking down the task of inference to smaller parts and are concerned with doing such "easy" inferences that are however essential for NLI.

5.1. Correcting a sub-corpus of SICK

One of the strengths of our approach is its precision with respect to a given lexicon. If some entailment is in the lexicon, it will be annotated correctly and the evidence of the entailment can be provided. But how close really are annotators' intuitions to the ones of the linguists that built lexical resources like PWN? Do they agree that a *a chef is cooking a meal* implies *a chef is preparing a meal*? Do they think that *typing* is *writing with a keyboard*? It seems that there is much disagreement as we already discussed in Kalouli et al. (2017a) and we could see once more in this work, something very astonishing if we take into account that these are "easy" inferences. Note that out of the 1651 pairs that PWN could label, 336 got a different label by the SICK annotators and accepting PWN as the correct, golden standard for such definitions, we can claim that 20% of this sub-corpus of SICK was wrongly annotated. Such a percentage raises worries, especially considering the fact that these are classified as "easy" inferences. So, the first contribution of our approach is to provide another corrected sub-corpus of SICK as we did before (Kalouli et al., 2017b) but this time with less effort. The corrected sub-corpus is available under `https://github.com/kkalouli/SICK-processing/corrected`.

5.2. Evaluating WordNet

Everyone should agree that there is an easy inference from the sentence *A dog is barking at a ball* to the sentence *An animal is barking at a ball*. Similarly no one would disagree with the assertion that *The baby elephant is not eating a small tree* contradicts the statement *The baby elephant is eating a small tree*. These are the kinds of trivial, noncontroversial inferences that SICK is expected to account for because its construction process was conceived exactly to add these kinds of inferences to sentences extracted from captions. But do our lexical resources support these trivial inferences? To what extent?

We were able to answer such questions by looking at our "one-word difference" approach and investigating which cases could be handled by WordNet and which ones are missing. We have provided the taxonomies in section 4. and these could be taken into account by lexicographers to improve PWN and other such resources. Some of the data presented above bring up old but interesting questions for further discussion, e.g. what is part of the definition of a noun (cf. example of *sewing machine*) and what is a relation of the word? We believe that the task of inference can and should be broken down to "easy" inferences like these ones and that therefore it is of great importance to have trustworthy, high-coverage resources that can solve big parts of them. Of course, lexical resources will never cover everything but they should always be expanded and then further supported by other state-of-the-art techniques such as word embeddings. We should also note at this point that our approach allows us to relatively evaluate the quality of the other tools used apart from PWN, i.e. the Enhanced Stanford Dependencies and the JIGSAW algorithm (cf. *amalgamating* example). It allows us to identify cases where PWN could not give a label not because of its own weakness but because the wrong sense was given to a word and this sense was not somehow related to the sense of the other word or a wrong part-of-speech was assigned which of course led to the wrong sense and thus again to no match.

5.3. "Easy" inferences as an evaluation standard

Evaluating lexical resources is a time-consuming task, mainly because we need to find appropriate test data which should on the one hand efficiently test the coverage of the resources themselves and on the other hand originate from real NLP scenarios that bring to light the whole complexity of language and thus the challenging cases. Our simple yet effective approach shows that tasks of "easy" inferences where the inference relation boils down to lexical relations that such lexical tools should account for, are a good method for testing and evaluating lexical resources. On the one hand, they offer concrete testing of the coverage because they can point out not only whether something is missing altogether but also if a word is missing some inter-relations essential for any NLP task (e.g. several adverbs *amusedly, amazedly, athletically* have only the adjective counterparts in PWN.). On the other hand, "easy" inferences that are extracted from corpora like SICK offer a reasonable and real-world testing scenario because it is exactly these corpora that are used for development or training of further NLP applications and it is thus important to test that the coverage for these corpora is there. Taking this suggestion a step further, we think that there should be an organized attempt to collect such real testing data from corpora and other similar language resources. No matter if these resources are inference-geared or not, it is important that "easy" inferences can be extracted from them so that testing data can be created. This means that it is important to be able to extract "one-word difference" sentences that can be used as testsuite data for the lexical tools. The more the lexical resources improve and expand with this method, the more complex the inference test cases should become so that we can reach a point where lexical resources are almost-complete, mature tools to deal with the first heavy lifting of reasoning.

Finally, we should remark that the approach here is very different from the one taken in the SemEval 2014 competition (Marelli et al., 2014a), where SICK was used as a testing corpus. Out of more than 20 participating teams in SemEval 2014, the top four performing systems are systems that build statistical classifiers on shallow features such as word alignments, syntactic structures and distributional similarities. Thus, our approach is incomparable to these, as we build a rule-based system that does not employ a statistical classifier at all and we only deal with one third of the corpus. The comparison with logic or hybrid logic-statistical systems is also hampered by the use of different grammatical and logical formalisms. We can suggest, following (Martínez-Gómez et al., 2017), that while we eventually envisage a system of Natural Logic or First-Order Logic, for the time being we only use the logic of PWN relations, which correspond to synonymy and subsumption between synsets, as well as simple heuristics.

6. Conclusions

We presented our PWN-based automatic approach for doing what we called "easy" inferences. With this approach we attained three goals: a) we could provide a corrected sub-corpus of SICK, b) we could evaluate facets of PWN and provide taxonomies of "easy" inferences and of PWN strengths and weaknesses and c) we observed that our approach is a suitable evaluation standard for lexical resources like PWN. We hope that this work can positively contribute to the improvement of WordNet which we would like to use further for our system of computing inference. We also hope that the concrete commenting and classifying of the PWN-labelled resource we provide publicly can raise interesting discussion points in the community. Last but not least, we believe that the suggestions coming from this work can be integrated in the general discussion about evaluating lexicographic resources and can help in future tasks. Continuing this work, we would like to expand and pursue further all our three goals. We would like to come up with additional ways of automatically correcting the SICK corpus or, at least, parts of it. Furthermore, we intend to try our method on other suitable inference corpora like SNLI in order to see if we can provide further PWN evaluation and additional "easy" inferences as test data for the evaluation of lexical resources. Finally, we would like to compare the inference relations and the taxonomies we rediscovered from WordNet and the ones suggested by the annotations, to other inference relations obtained by researchers interested in precision focused inference over SICK such as (Beltagy et al., 2015).

7. Bibliographical References

Basile, P., de Gemmis, M., Gentile, A. L., Lops, P., and Semeraro, G. (2007). UNIBA: JIGSAW algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic, June. Association for Computational Linguistics.

Beltagy, I., Roller, S., Cheng, P., Erk, K., and Mooney, R. J. (2015). Representing meaning with a combination of logical form and vectors. *arXiv preprint arXiv:1505.06816 [cs.CL]*.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

de Paiva, V., Bobrow, D. G., Condoravdi, C., Crouch, D., Nairn, R., and Zaenen, A. (2007). Textual inference logic: Take two. In *Proceedings of the International Workshop on Contexts and Ontologies: Representation and Reasoning (C&O:RR) Collocated with the 6th International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-2007), Roskilde, Denmark.*

de Paiva, V., Real, L., Oliveira, H. G., Rademaker, A., Freitas, C., and Simões, A. (2016). An overview of Portuguese WordNets. In *Global Wordnet Conference 2016*, Bucharest, Romenia, January.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Hoeksema, J. (1986). *Monotonicity phenomena in natural language*. Linguistic Analysis.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017a). Correcting contradictions. In *Proceedings of Computing Natural Language Inference (CONLI) Workshop, 19 September 2017*.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017b). Textual inference: getting logic from humans. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014a). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014b). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

Martínez-Gómez, P., Mineshima, K., Miyao, Y., and Bekki, D. (2017). On-demand injection of lexical knowledge for recognising textual entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 710–720.

McCartney, B. (2009). *Natural Language Inference*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI3364139.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Chris Welty et al., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.

Pavlick, E. and Callison-Burch, C. (2016). Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany, August. Association for Computational Linguistics.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.