Data Creation Guidelines for the SemEval 2010 Task 8

Multi-Way Classification of Semantic Relations Between Pairs of Nominals

http://docs.google.com/View?docid=dfvxd49s_36c28v9pmw

1. Annotation Objective

Our objective is to annotate instances of semantic relations that are **true** in the sense of holding in the most plausible truth-conditional interpretation of the sentence. This is in the tradition of the SemEval 2007 task #4 on Semantic Relations, and the Textual Entailment/Information Validation paradigm, and a goal distinct from "aboutness" annotation such as provided by Semantic Role Annotation or the BioNLP 2009 task.

A range of structural and lexical factors affect the truth value, in particular negation (both overt and lexical, e.g. in the form of embedding verbs) and modality (introduced e.g. by modal verbs and particles). Also, in our experience, the truth status of a semantic relation between two words becomes considerably more debatable (defeasible) the more distant the words in the syntactic structure of the sentences. To counteract these tendencies, this document lists various constraints on example sentences that are to be applied during the construction of datasets for relations. We aim to make these constraints as mechanically applicable as possible, although not at the cost of excessive overhead.

By necessity, the criteria in this document are incomplete. They aim to establish a "lower bound" -- sentences that violate these constraints should not be annotated. However, not all sentences that conform to the constraints are necessarily appropriate examples.

All nine relations which we currently consider are asymmetric, so that Relation(X, Y) => not Relation(Y, X).

2. Locality

The two nominal expressions should be "local" to one another. We include the following constructions, which we consider as clear cases of "local" relationships:

- Simple clauses: *The <e1>glass</e1> contains <e2>beer</e2>.*
- Relative clauses: *A <e1>glass</e1> that contains <e2>beer</e2>* / *A <e1>glass</e1> containing <e2>beer</e2>*
- Compound nouns/base NPs: *A <e1>beer</e1> <e2>glass*</e2> (whether the glass actually contains beer is not a locality issue)
- Modifiers in simple clauses: *When <e1>philosophy</e1> is the topic of <e2>discussion</e2>...*

This list may be amended.

We also aim to exclude sentences which are syntactically overly complex (extraposition, cleft constructions etc.) because syntactic variability is not our focus. We do, however, allow adjuncts between e1 and e2, such as non-restrictive relative clauses or adverbial phrases attached to the NP containing e1. Two examples:

(with an adjunct)
Four sets of wax<e1>glands</e1>, situated inside the last four ventral segments of the abdomen, produce<e2>wax</e2> for comb construction.

(without an adjunct)
Four sets of wax<e1>glands</e1> produce<e2>wax</e2> for comb construction.

(with an adjunct)
The <e1>connection</e1> between testing and the workplace originated early in the last <e2>century</e2>.

(without an adjunct)
The <e1>connection</e1> originated early in the last <e2>century</e2>.

3. "Real World" Situations

We consider only instances of semantic relations as pertaining to situations in the "real world" and we exclude instances as pertaining to situations in some other world defined by counterfactual constraints elsewhere in the context. For example, in the sentence:

"*Suppose you were given a bottle that contains 400 grams of a 3.0% bleach solution*."

The presence of the "bleach solution" inside the "bottle" is a situation being described as holding in a counterfactual hypothetical world.

We therefore exclude top-level clauses with

- overt negations (e.g. *not*)
- modal verbs (e.g. *can*, *may*, *shall*, *would*)
- modal adverbials (e.g. *maybe*, *possibly*, *probably*, *certainly*)
- opinion verbs (e.g. *think*, *believe*, *suppose*)

which take scope over at least one of the nominal expressions in question, or the complete clause. We also exclude conditional clauses (if, unless, assuming that...) and imperative clauses.

We consider examples involving motion verbs (e.g. "put", "remove", "run", "enter", etc.) -- that is, verbs actually describing a movement activity -- as Entity-Destination or Entity-Origin examples, according to the direction of the motion.For example:

"*I put/removed the <e1>apples</e1> in/from the <e2>basket</e2>*."

"*I've put/removed the <e1>apples</e1> in/from the <e2>basket</e2>*."

"*I'm putting/removing the <e1>apples</e1> in/from the <e2>basket</e2>*."

"*I was putting/removing the <e1>apples</e1> in/from the <e2>basket</e2>*."

"*The <e1>girl<e1> ran away from her <e2>family</e2>*."

"*One basic trick involves a spectator choosing a <e1>card</e1> from the <e2>deck</e2> and returning it*"

When using motion verbs, the lexical choice, perspective focus, and emphasis is on the movement relation and prevail over "stative" relations such as Content-Container, Component-Whole, or Member Collection. Note that such "stative" relations might be inferred with a certain confidence and describe a possible eventual outcome, or being actually true (the girl is anyway a member of her family) but we don't care.

A scenario recalling possible eventual outcomes can be found even in some progressive sentences. Let us consider:

1. "*That man was/is building his home doing all his own labor*. "
2. "*That man built his home doing all his own labour*."

Both the sentences refer to the process of building a house. But while 2. refers to a process which leads to a culmination, 1. does not assert that the culmination of the process occurred. Process 1 would lead to the culmination only if it were to continue uninterrupted. For this reason we do not consider 1 as a positive example of Product-Producer, while 2 is positive. Progressive sentences do not all denote a non complete process:

The factory is producing 50,000 cars per day.

The progressive here denotes iterativity of the process rather than an ongoing process. So this actually is a positive example for the Product-Producer relation.

4. Independence from Discourse

We do not annotate examples whose interpretation relies on discourse knowledge. As a simple indicator, we exclude sentences where the entities that we annotate consists of anaphoric expressions such as personal and demonstrative pronouns referring to sentence-external material.

5. Restrictions on Nominal Expressions

We consider as markable only base noun phrases whose head is a common noun. Named entities, whose behavior differs considerably from common nouns, are excluded. A base noun phrase (Base NP) is a noun and its premodifiers (e.g., nouns, adjectives, determiners). We do not include complex noun phrases (e.g., noun phrases with attached prepositional phrases). For example, "lawn" is a noun, "lawn mower" is a base noun phrase, and "the engine of the lawn mower" is a complex noun phrase.

For reference, in Winograd's (1982) definition of NPs, a Base NPs would be anything that stretches from Pre-determiner to Head (looser reading) or from Describers to Head (stricter reading) as shown in the table below. We opt for the stricter reading.

Segment	**Function**	**Examples**
Determiner	Pre-determiner	half; both; all
sequence	Determiner	the; a; this; every
	Ordinal	first; second; last
	Cardinal	one; two; three
Modifiers	Describers	big; purple; enchanted
	Classifiers	
Head	Head	
Qualifiers	Restrictive q.	in town; that flies
	Nonrestrictive q.	John, whom you know
--------	Possessive marker	-'s

Given the NP "*a brown dwarf star*", there are five segments which have the structure of base NPs:

dwarf -- star -- dwarf star -- brown dwarf -- brown dwarf star

In our annotations, the entities e1 and e2 will typically span a single word; they can only span two or more words in case of partial or full lexicalizations (see www.merriam-webster.com/dictionary/lexicalization for a definition), e.g., we should have "<e1>science fiction</e1> <e2>writer</e2>" rather than "science <e1>fiction</e1> <e2>writer</e2>" since the noun compound "science fiction" is more than a special kind of "fiction". Similarly, we cannot remove "health" from e1 in "<e1>health care</e1> </e2>provider</e2>".

6. Sense Ambiguity

We do not annotate the entities with WordNet senses (unlike in SemEval-2007 Task 4). This makes the task more realistic but it makes the task of the manual annotators harder. Some examples have ambiguous entities in which it is difficult to determine which sense applies, multiple senses could be true at the same time. Here is an example of the ambiguous entity '*village*'. For this case annotators disagree with each other and there is no most plausible interpretation of the sentence. Such examples are discarded from the data set.

"In that valley you'll find a lot of this old and historical <e1>villages</e1> with beautiful <e2>churches</e2>."

Component-Whole(e2, e1)

A village is an organizational unit and a church is an integral and functional part of a village.

Or:

Entity-Location(e2, e1)

A village is a geographical area and a church is located in the village.

7. Noun Compounds

Noun-noun compounds are compressed propositions usually conveying a relationship between the two concepts that are expressed by the two nouns. We annotate such compounds as positive for a relation even when the context is ambiguous provided that the relation is understood to be generically true. For example, we annotate "Bees produce honey." as positive for Product-Producer even though the sentence is not referring to a particular state of affairs. Similarly, we accept the noun compound "honey bee" as a potential Product-Producer that basically means "bee that produces honey". On the other hand, "wine bottle" can be paraphrased as "bottle that contains wine" (Content-Container), but also as "a bottle for wine" (Purpose-Tool) -- in that case, we cannot make a generic choice out of context. Thus, "There is a wine bottle on the table." would be Other since the context does not allow us to decide between the two readings and there is no generically true out-of-context reading.

8. The Passage of Time

The goal of our relation definitions was to reduce *vagueness* as much as possible: that is, we wanted to reduce the "grey areas" between relations. However, some instances (which could be called "conjunctive cases"), remain problematic, because the issue is not that they lie *between* relations A and B, but rather that they invoke *both* A and B. This happens because events in real life tend to follow one another in the course of time -- and a Product X that is produced by a Producer Y, for example, tends to move away from the Producer, and might thus give rise to the relation Entity-Origin(X,Y). See the following example:

<e1>Hamburgers</e1> from the <e2>restaurant</e2> were recalled, preventing further illness."

In order to assign a single relation to similar cases nevertheless, we use the following heuristics:

1. We do not assign relations when the sentence states specifically they do not hold at present any more (i.e., when they only held at some earlier point in time).
2. Example: even though "The <e1>ball</e1> is retrieved from the <e2>hole</e2>." might evoke both Content-Container and Entity-Origin, Content-Container is out because the ball has been removed from the hole.
3. We apply the motion verbs prevailing over "stative" relations criteria described in Section 3.
4. We use the following rough "informativity order" (less informative < more informative) and prefer more over less informative relations:

   Entity-Origin / Entity-Destination < Message-Topic < Instrument-Agency < Content-Container

   < Component-Whole / Member-Collection < Cause-Effect /Product-Producer

   so that the Hamburgers/restaurant example from above would be Product-Producer rather than Entity-Origin.