Computational Semantic Detection of Information Overlap in Text

Julia M. Taylor (jtaylor@riverglassinc.com)
RiverGlass Inc, 2001 South First St
Champaign, IL 61820 USA

Abstract

This paper is an attempt to investigate whether a computer is capable of finding similar information in structurally different texts, as people do it, without relying on lexical matching and without guessing the meaning of sentences based on word co-occurrence. Considered texts describe the same event, but each text may focus on different parts of the event. The considered texts are not paraphrases, but rather human-produced descriptions of a simple picture. The goal is not to find similar words in texts, which can be easily done, but to meaningfully connect the overlapping concepts and relationships used in the text descriptions. The meaning-based approach does not use any statistical/machine-learning techniques. The performance of a machine in finding similarity is compared to human performance not just in numbers but in the found information. The results show that the machine matches four out of the five human findings.

Keywords: text duplication and similarity, information overlap detection, meaning processing, ontological semantics.

Overview

This paper examines the use of the Ontological Semantic Technology (OST)—a modified version (Raskin et al 2010) of Ontological Semantics (Nirenburg & Raskin 2004)—for processing similar texts and compares it to human processing. Instead of selecting existing texts and assessing their similarity, users were given the same picture to describe. Clearly, the users will emphasize different objects or events on the picture, but at the same time, because they are all looking at the same picture, some of the provided information will overlap. The experiment is done to demonstrate the ability of the technology to understand the meaning of text, regardless of individual words that are used and of the length of the sentences.

The OST claim to fame is that it "understands" the meaning of text. The meaning of text includes paraphrases of sentences or paragraphs. A large number of paraphrases can be produced from a single sentence, an even larger number can be produced from a paragraph. Because of this large number of potential paraphrases, and because it is unclear which ones are good enough, instead of asking people to paraphrase a text, we ask them to describe a picture.

The untested assumption is that looking at the picture should activate the same schema(ta) as reading a paragraph. Thus, the main information received should be approximately the same whether looking at the picture or reading text. Instead of *reconstructing* original sentences after reading or listening to a text, the subjects were asked to *describe* what they see on the picture in their own words. The tasks of paraphrase and describing a picture are by no means identical, even for short sentences when compared to very simple pictures. Several things should be noticed: 1. Length of the sentences in paraphrases has probably some correlation to the length of the original sentences. 2. The choice of words for the description task is not limited by the original sentence, whereas it is possible that, in the paraphrase, the subjects would try to come up with unnatural synonyms in their desire to paraphrase. 3. The order of sentences is free in the picture description, while it is possible that the sentences would be ordered according to the original text in the paraphrase.

While paraphrase detections have received some attention from the machine-learning community (Fernando & Stevenson 2008, Clough et al 2002, Qiu et al 2006, Zhang & Patrick 2005), to the best of our knowledge same picture descriptions have not been addressed. This is surprising because most real life event descriptions are more similar to picture descriptions than to paraphrasing tasks.

The task of paraphrase limits information that is available to the subjects to that in the task, while describing the picture provides more freedom of focus. For example, the sentence *a black ball is on top of a green cube*, can only be paraphrased in term of the provided information. Possible paraphrases are: *a green cube is under a black ball; a black sphere-shaped object is above a green cube; a ball is positioned on top of a cube, the ball is black and the cube is green*. Notice that there may be a considerable variation among paraphrases in terms on the words used, the order in which they are described, and the number of clauses used in the description. What they all have in common, however, is the properties and attributes that connect the described objects: all describe shape either explicitly as in *sphere-shaped object*, or by accessing the knowledge of a shape of a lexical item as in *ball* or *cube*; and all describe color. However, if picture is shown (Figure 1), other things may come into focus for different people, such as relative size of the objects.

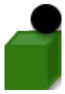

Figure 1: A black ball on top of green cube

It would be interesting to see if such unmentioned-in-the text characteristics would ever be brought up by the subjects in the paraphrase generation as unknown. It is, however, not the purpose of this experiment. The only significant assumption for this paper is that the greater variation of text should be encountered in the picture experiment, which in turn tests the machine's capability of catching the overlap to a much greater extent. On the other hand, it would be interesting to see if a coherent description of a situation could be constructed from a union of all descriptions, as it is likely that these descriptions, to some extent, complement each other.

It is this overlap information in descriptions reported by subjects, as well as the difference or the union, that is captured and analyzed by the machine, as compared to the overlap and difference in information in responses as perceived by human is the subject of the paper. The theoretical knowledge obtained in this kind of research is applicable to an increasingly urgent task of easing the information overload by removing duplicate and overlapping information[1].

Ontological Semantics Technology

OST is an upgraded, much improved and implemented (and, on occasion, perverted) version of (Nirenburg & Raskin 2004) that detects the meaning of text. Ontological Semantics is a theory, methodology and technology for representing natural language meaning, for automatic transposition of text into the formatted text-meaning representation (TMR), and for further manipulation of TMRs for inferencing and more advanced reasoning, both theoretically and in a growing variety of applications. The main knowledge resources in OST are the language-independent ontology and language-specific lexicons.

The OST is not a toy system that works on a handful of examples; instead, it works with unrestricted texts in real-life applications, as well as avoiding the scalability problems (see Raskin et al 2010).

Ontology

The ontology contains information about the world; it is a constructed, engineered model of reality, a theory of the world (Gruber 1993, 1995; Nirenburg & Raskin 2004:138-139). It is a structured system of concepts covering the processes, objects, and properties in all of their pertinent complex relations, to the grain size determined by an application or by considerations of computational complexity. The ontology contains PROPERTIES, EVENTS, and OBJECTS. The concepts are named purely for the convenience of a human: the label itself does not contribute to the information content. Every OBJECT and EVENT is defined with a number of properties, thus allowing the concept to differ not only in label, but also in machine-understandable information. The child concepts inherit properties from the parent concepts.

Formally, the OST ontology is a lattice of conceptual nodes (for a construction of ontology and verification see Hempelmann et al. 2010 and Taylor et al 2010 respectably), each of which is represented as:

concept-name
(property (facet(property-filler$^+$))$^+$)$^+$
property-filler
concept-name | literal value
property
attribute | relation
facet
SEM | VALUE | DEFAULT | RELAXABLE-TO[2]

The current implementation of OST uses the following three axioms:

- subClassOf for concepts: IS-A (example: PHYSICAL-OBJECT IS-A OBJECT)
- subPropertyOf for properties: IS-A (example: COLOR IS-A PHYSICAL-OBJECT-ATTRIBUTE)
- inverse for properties: INVERSE (example: THEME INVERSE THEME-OF)

Concept interpretation (without facets, for the ease of reading) can be looked at using the following: given a set of objects $\mathcal{D}$, where $\mathcal{D}$ is the disjoint union of $\mathcal{D}$c (concepts) and $\mathcal{D}$d (literals), and given its interpretation function $\mathcal{I}$, for every atomic concept B, $\mathcal{I}$[B]$\subseteq\mathcal{D}$c; for every literal V, $\mathcal{I}$[V]$\subseteq\mathcal{D}$d ; for every relation R, $\mathcal{I}$[R]$\subseteq\mathcal{D}$c x $\mathcal{D}$c; for every attribute A; $\mathcal{I}$[A]$\subseteq\mathcal{D}$c x $\mathcal{D}$d. Moreover, the following is true for concepts C and D:

$\mathcal{I}[\text{ALL}] = \mathcal{D}$
$\mathcal{I}[\varepsilon] = \emptyset$
$\mathcal{I}[\text{C D}] = \mathcal{I}[\text{C}] \cup \mathcal{I}[\text{D}]$
$\mathcal{I}[\text{and C D}] = \mathcal{I}[\text{C}] \cap \mathcal{I}[\text{D}]$
$\mathcal{I}[(\text{Rel(D)}))]= \{x\in\mathcal{D}\text{c}|\ y \in \mathcal{I}[\text{D}], <x, y>\in\mathcal{I}[\text{Rel}]\}$
$\mathcal{I}[(\text{Rel(and C D)})] = \mathcal{I}[\text{Rel(C)}] \cap \mathcal{I}[\text{Rel(D)}]$
$\mathcal{I}[\text{Rel(C D)}] = \mathcal{I}[\text{Rel(C)}] \cup \mathcal{I}[\text{Rel(D)}]$
$\mathcal{I}[\text{C(Rel(D))}] = \mathcal{I}[\text{C}] \cap \mathcal{I}[\text{Rel(D)}]$
$\mathcal{I}[(\text{Att(V)}))]= \{x\in\mathcal{D}\text{c}|\ y \in \mathcal{I}[\text{V}], <x, y>\in\mathcal{I}[\text{Att}]\}$

Clearly, concept C is a descendant of D if $\mathcal{I}$[C]$\subseteq\mathcal{I}$[D]; and $\mathcal{I}$[(C(R(D))]$\subseteq\mathcal{I}$[C]. Whenever relation Rel is defined with a domain D and range R, if $\mathcal{I}$[C]$\subseteq\mathcal{I}$[D] and $\mathcal{I}$[E]$\subseteq\mathcal{I}$[R], then C(Rel(E)) is equivalent to $\mathcal{I}$[C] $\cap\mathcal{I}$[D(Rel(and E R))].

For the examples in this paper, it is sufficient to mention that when facets are involved, the highest priority facet takes precedence over the lower priority one.

Lexicon

The lexicon is the starting point for machine interpretation of language in OST. Since Ontological Semantics is centered on meaning, we will largely concentrate on the semantic structure (sem-struc) part of the lexicon entries.

In general, the lexicon can be looked at as a collection of words (and phrasals), organized such that each word is

[1] The author believes that whether an overlap indicates an importance of information in text is a separate (to her, dubious) hypothesis, which will not be addressed in this paper.

[2] The list shown has been enriched in the current implementation of OST, but since facets do not contribute much to this paper, the list is left as it was first introduced.

listed with all of its senses. Each sense of the word in a lexicon follows the following structure:

```
(WS-PosNo
(cat(Pos))
(synonyms "WS-PosNo"))
        (anno(def "Str")(ex "Str")(comments "Str"))
        (syn-struc((M)(root($var0))(cat(Con))(M))
        (sem-struc(Sem))
)
```

where the following grammar defines what is allowed:

```
M  → (Srole((root(Var))(cat(Cpos)))
   → (Srole((opt(+))(root(Var))(cat(Cpos)))
   → (M(M))
Pos → N |              (noun)
    → V |              (verb)
    → Adj |            (adjective)
    → Adv |            (adverb)
    →…
Con → NP |             (as defined by rules omitted
    → VP |             here to save space)
    → Con Con |
    → Pos
SRole  → subject |            (syntactic roles,
       → directobject |       only some are shown
       → pp-adjunct           to save space)
       → …
No     → [1-9]                (any digit)
Str    → [A-Z|a-z| |,|.]      (any string)
Var    → $varNo
       → Str
Sem    → C |                  (any ontology concept)
       → ^Var(R(F(C))) |      (R, F, C from ontology)
       → C(R(F(^Var)))        (C, R, F from ontology)
```

When the machine processes text with the help of the resources, the ontological concepts are accessed through the (English) lexicon. For example, a lexical entry for the verb *run* will contain all the possible senses, of which #6 is shown below:

```
(run-v6
    (cat(v))
    (anno
        (comments "...")
        (def "meet unexpectedly")
        (ex "I ran into my teacher at the movies last
night."))
    (syn-struc
        ((subject((root($var1))(cat(np))))
        (root($var0))(cat(v))
        (prep((root(into))(cat(prep))))
        (directobject((root($var2))(cat(np)))))
    )
    (sem-struc
        (meet-with
            (agent(value(^$var1(should-be-
a(sem(human))))))
            (beneficiary(value(^$var2)))
            (intentionality(value(<0.3))(relaxable-to(<0.5)))
        )
    )
)
```

The entry shows that this sense of *run* means ‘unexpected meeting event’ (from sem-struc), and it needs a preposition *into* (from syn-struc) to be activated. It also shows that in its normalized form the subject is usually the agent of the event, and the direct object is the beneficiary. Optional properties such as time, place, etc are usually not shown in the lexical items.

OST On Black balls and Green Cubes

OST uses the Semantic Text Analyzer (STAn) to interpret the meaning of sentences. The (machine generated) output of STAn is a text meaning representation (TMR) that shows the conceptual representation of the text, regardless of the language of the input. Let us go back to the sentence *a black ball is on top of a green cube*. The resulting TMR is:

```
Event: pred1
  (theme(value (physical-object1
    (shape(value(sphere)))
    (color(value(black)))
    (above(value(physical-object2
        (shape(value(cube)))
         (color(value (green)))
    )))
)))
```

Possible paraphrases from the previous section is: *a green cube is under a black ball:*

```
pred1
  (theme(value (physical-object1
    (shape(value(cube)))
    (color(value(green)))
    (below(value(physical-object2
        (shape(value(sphere)))
         (color(value (black)))
    )))
)))
```

Another interesting paraphrase is: *a ball is positioned on top of a cube, the ball is black and the cube is green*, which will result in the following:

```
put1
  (theme(value (physical-object1
    (shape(value(sphere)))
    (above(value(physical-object2
        (shape(value(cube)))
    )))
)))
pred1
   (theme(value (physical-object1
        (shape(value(sphere)))
        (color(value(black)))
    )))
  pred2
    (theme(value (physical-object2
         (shape(value(cube)))
         (color(value(green)))
    )))
```

Notice that besides the PUT event, corresponding to *is positioned*, and the inverse of the BELOW-ABOVE properties, the rest of the information is identical for any purposes, including reasoning. The third example is especially interesting, as the colors are assigned to the indexed objects, referenced by the previous sentence.

The intersection of the paraphrases, as indicated by the TMRs once the inverse properties are used, are:

```
   pred1
     (theme(value (physical-object1
       (shape(value(sphere)))
       (color(value(black)))
       (above(value(physical-object2
            (shape(value(cube)))
             (color(value (green)))
       )))
  )))
```

The union of the TMRs adds information only present in the third example, namely that of PUT, thus, producing

```
  put1
     (theme(value (physical-object1
       (shape(value(sphere)))
       (color(value(black)))
       (above(value(physical-object2
            (shape(value(cube)))
             (color(value (green)))
       )))
  )))
```

If Figure 1 is described instead of paraphrases, and sentences like a *ball is smaller than a cube* happen to be added to the description, it is easy to see that the intersection of TMRs will remain the same, while the union will add the additional size information.

More Complex Pictures

As demonstrated in the previous sections, OST is capable of understanding the meaning of close paraphrases and represent it in such a way that the differences and similarities are shown. The next experiment aimed at stretching the similarities as far as possible, but asking the user to describe a picture instead of paraphrasing a text.

The picture shown to the user was selected to depict an unambiguous object in the foreground, while the background contains objects that can be described either very briefly, if at all, or be paid as much attention as possible. The hypotheses are:

- The description of the central element of the picture is affected by individual/personal schemata, and therefore will partially differ from person to person. However, there should be an overlap in descriptions, focused on that central object, just as the paraphrases showed.
- The description of the background will differ from person to person to a much greater degree. A very small overlap is expected from pairs of participants since the background is not in focus (metaphorically).
- The activated schemata are not expected to be known to a computer, thus the computer will process only information explicitly stated by the subjects.

This is not at all an attempt to deal with the well-researched figure-ground phenomenon (see Talmy 2000, vol, 1: 311-344). Instead, we are only interested in the foreground display, but the background may provide individual distinctions.

Methodology

Once a picture was chosen, 3 subjects, unfamiliar with an experiment's goals and from unrelated occupations, were asked to describe the picture. The picture was visible to the subject all the time, thus the description is not effected by the accuracy of their recollection of the picture. The instructions requested to describe only what is seen on the picture, without alluding to any inferences or encyclopedic knowledge that the picture may activate. The subjects were not given any specific time frame to complete the task.

The described text was then entered into a machine for processing, and the union and intersection of information in individual texts were computed. Whenever the descriptions contradicted each other, the contradictions were also added to the union as alternative interpretation.

To check the validity of the found union and intersection, a person not participating in the description task and not involved in the OST part of the experimentation was asked to highlight the similarities in text. These similarities were then compared to the intersection of interpretations provided by a computer.

The foreground of a picture showed a moving elephant. The background of the picture contained trees, shrubs and other greenery, as well as a place where several cars were parked, as seen in Figure 2.

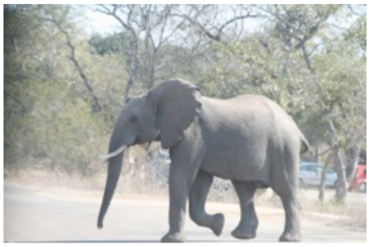

Figure 2: An elephant crossing the road

Results of Human Description

The descriptions of the submitted texts varied length (the first text used 54 words, the second text used 124 words, and the third text used 151 words) and structure of sentences.

The following similarities were noticed by a human in all of the descriptions:

- Elephant's existence.
- Road on which the elephant is located.
- Trees in front of the cars, in some spatial relation to the elephant
- Cars parked in the background

The following information was included in at least one of the texts (author's summary below):

- A large African male elephant is shown on the picture and is moving either on the road or bare ground or crossing the road. The elephant has large tusks, 4 legs, one visible ear, one visible eye, a tail and a trunk. The front right leg of the elephant is bent at the knee.
- There is dust on road and some dirt or hard soil on the edges of the road. The road is wide and paved.
- A row of trees are between the elephant and the cars, past the cars and on the berm. The trees are large with extensive but not overwhelming foliage. The grass is mostly yellowish and dusty.
- Cars, red and light blue or white, are parked on the parking lot. The red car is a hatchback. The cars, either 4 or 2, are all compact models. All cars are parked behind the trees on what may be a parking lot.
- A building that has yellow corner is behind the cars.
- It is a bright sunny day; the sky is blue with light clouds.

From this description, it can be noticed that the hypothesis of the central element of the picture being similarly described between all participants could not be accepted. Interestingly, the descriptions varied in movement information—it could be argued that it is not salient to the central object itself—but not in the elephant's location on the road. The description of the elephant and its body parts did not vary as much between any 2 subjects as between all of them. It should also be noticed that there was no contradictory description of the elephant itself. Thus, perhaps a better metric would be to find overlap used by the majority of the participants, as opposed to all, for real-world applications.

The second hypothesis, namely the difference in the background descriptions due to focus on different elements could not be rejected based on this small set. Between the objects that were noticed by all participants, the description varied more than that of the central object, and often the information was contradictory. For example, there was no agreement on the number of cars in the picture or their colors and very different description of greenery.

Computational Description

Computational overlap, as expected, was clustered around objects. Thus, the following concepts were identified: ELEPHANT, ROAD, CAR, TREE. Additionally, the following descriptions of the concepts were found:

```
undetermined_event
    (agent(value(elephant1)))
    (location(value(road1)))
car1
    (behind(value(tree1(number(greater-than(1))))))
put2
    (instrument(value(car1)))
    (location(sem(parking-lot)))
```

In plain English, it says that there is an elephant that is doing something on the road, there is a car behind trees, and somebody left a car in the parking-lot. Clearly, what is missing here from the overlap found by a human is that there are trees in some special relation to the elephant.

The union of information was not as successful due to coreference resolution mistakes (with STAn's coreference module not yet fully activated), however, the trivial unions of information were found. The number of unconnected clusters of information was small enough, that based on the concepts connected through the overlap above, it is possible to conclude that the three stories described similar information.

Perhaps it is worthwhile to demonstrate the computational process in the discovery of the overlap. Consider the following sentences:

- (1) A large grey elephant is moving on a road or bare ground.
- (2) This is a photograph of an elephant crossing a road. It is a large male African elephant.
- (3) Elephant is on asphalted road.

The sentences result in the following TMRs:

```
(1) land-animal-motion1
      (phase(value(continue)))
      (agent(value (elephant1)))
         (color(value(grey)))
      )))
      (location(value(road1 ground1)))
(2) pred1
      (theme(value(photograph
            (representation-of(value(change-location1
                  (agent(value(elephant1)))
                  (path(value(road1)))
            )))
      )))
    pred2
      (theme(value(elephant1
            (size(value(large)))
            (gender(value(male)))
            (location(pnd(Africa)))
      )))
(3) exist1
      (agent(value(elephant1)))
      (location(value(road1
            (made-of(value(asphalt)))
```

)))

From the above descriptions, we know the following about the elephant:

From (1): <land-animal-motion1, elephant1> $\in I$ [agent]
From (2): <change-location1, elephant1> $\in I$ [agent]
From (3): <exist1, elephant1> $\in I$ [agent]

Taking the intersection of the events for which the elephant is an agent results in x $\in I$ [event]. Thus, producing undetermined_event(agent(value(elephant1))).

Continuing with each TMR, we find the following:

From (1): <land-animal-motion1, road1> $\in I$ [location]
From (1): <land-animal-motion1, ground1> $\in I$ [location]
From (2): <change-location1, road1> $\in I$ [path]
From (3): <exist1, road1> $\in I$ [location]

It can be easily noticed that ground1 occurs only in (1), thus the intersection with (2) and (3) results in an empty set. For road1, the calculation is similar to that of an elephant with the only addition of parent-child relationship of location and path.

It should also be noted that if we were to find an overlap of (1) and (2) and discarded (3), the event in question would have a considerably finer grain. According to the ontology, the most specific ancestor of both LAND-ANIMAL-MOTION and CHANGE-LOCATION is CHANGE-LOCATION. This means that while the sentences used different verbs to describe the movement of the elephant (crossing and moving), the OST understands what both mean and finds the general concept for both, as opposed to ignoring the similarity in meaning.

Similar processing is done for all sentences, resulting in the above relationship for car1 and put2 in addition to elephant.

The calculation of overlap is done in a similar manner, with the exception of the selection rules: each pair of concepts does not have to overlap in the found properties, instead uniquely found relationships are added to the existing set.

Conclusion

This paper was an attempt to investigate whether a computer is capable of finding similar information in structurally different texts that describe the same event, each focusing on potentially different parts of the event. The goal was not to find similar words in texts, which can be easily done, but to meaningfully connect the overlapping concepts and relationships used in the text descriptions. The approach is radically different from the machine-learning one. The performance of a machine in finding similarity was compared to human performance. The machine matched four out of five human findings.

It is too early to reach a conclusion that it is possible for computers to find overlap and difference between texts similarly to those that humans find, and, of course, more extensive experiments should be conducted. However, it is promising that the first result is not negative.

Acknowledgements

The author is grateful to Victor Raskin and the anonymous reviewers for their comments and to RiverGlass Inc for permission to use examples from their proprietary resources.

References

Carroll, D. (2004), *Psychology of Language*, Thompson Wadsworth, Belmont, California, 2004

Clough, P., Gaizauskas, R., Piao, S. & Wilks, Y. (2002) METER: MEasuring TExt Reuse. In Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02), pages 152–159, Pennsylvania, PA.

Fernando, S. & Stevenson, M. (2008) A semantic approach to paraphrase identification. In *Proceedings of the 11th Annual Research Colloquium of the UK Special-interest group for Computational Lingusitics*, Oxford, England.

Gruber, T. R. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition*, 5, 199-200

Gruber, T. R. (1995) Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, (Eds.), Special Issue: The Role of Formal Ontology in the Information Technology. *International Journal of Human and Computer Studies* 43(5-6), 907-928

Hempelmann, C.F, Taylor, J. M., & Raskin, V. (2010) Application-guided Ontological Engineering, In *Proceedings of International Conference on Artificial Intelligence,* Las Vegas, Nevada

Nirenburg S., & Raskin, V. (2004) *Ontological Semantics.* Cambridge, MA: MIT Press

Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010) Guessing vs. Knowing: The Two Approaches to Semantics in Natural Language Processing, In *Proceeding of Annual International Conference Dialogue 2010*, Moscow, Russia

Qiu, L., Kan, M.Y, & Chua, T. S. (2006) Paraphrase recognition via dissimilarity significance classification. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 18–26, Sydney, Australia, July. Association for Computational Linguistics.

Talmy, L. 2000. Toward a cognitive semantics, vols. 1-2. Cambridge, MA: MIT Press

Taylor, J. M., Hempelmann, C. F., & Raskin, V. (2010) On an Automatic Acquisition Toolbox for Ontologies and Lexicons in Ontological Semantics, In *Proceedings of International Conference on Artificial Intelligence,* Las Vegas, Nevada

Zhang, Y. & Patrick, J. (2005) Paraphrase identification by text canonicalization. In Proceedings of the Australasian Language Technology Workshop 2005, pages 160–166, Sydney, Australia, December.