Automatic Acquisition of Ranked Qualia Structures from the Web[1]

Philipp Cimiano
Inst. AIFB, University of Karlsruhe
Englerstr. 11, D-76131 Karlsruhe
cimiano@aifb.uni-karlsruhe.de

Johanna Wenderoth
Inst. AIFB, University of Karlsruhe
Englerstr. 11, D-76131 Karlsruhe
jowenderoth@googlemail.com

Abstract

This paper presents an approach for the automatic acquisition of qualia structures for nouns from the Web and thus opens the possibility to explore the impact of qualia structures for natural language processing at a larger scale. The approach builds on earlier work based on the idea of matching specific lexico-syntactic patterns conveying a certain semantic relation on the World Wide Web using standard search engines. In our approach, the qualia elements are actually ranked for each qualia role with respect to some measure. The specific contribution of the paper lies in the extensive analysis and quantitative comparison of different measures for ranking the qualia elements. Further, for the first time, we present a quantitative evaluation of such an approach for learning qualia structures with respect to a handcrafted gold standard.

1 Introduction

Qualia structures have been originally introduced by (Pustejovsky, 1991) and are used for a variety of purposes in natural language processing (NLP), such as for the analysis of compounds (Johnston and Busa, 1996) as well as co-composition and coercion (Pustejovsky, 1991), but also for bridging reference resolution (Bos et al., 1995). Further, it has also been argued that qualia structures and lexical semantic relations in general have applications in information retrieval (Voorhees, 1994; Pustejovsky et al., 1993). One major bottleneck however is that currently qualia structures need to be created by hand, which is probably also the reason why there are almost no practical NLP systems using qualia structures, but a lot of systems relying on publicly available resources such as WordNet (Fellbaum, 1998) or FrameNet (Baker et al., 1998) as source of lexical/world knowledge. The work described in this paper addresses this issue and presents an approach to automatically learning qualia structures for nouns from the Web. The approach is inspired in recent work on using the Web to identify instances of a relation of interest such as in (Markert et al., 2003) and (Etzioni et al., 2005). These approaches rely on a combination of the usage of lexico-syntactic pattens conveying a certain relation of interest as described in (Hearst, 1992) with the idea of using the web as a big corpus (cf. (Kilgariff and Grefenstette, 2003)).
Our approach directly builds on our previous work (Cimiano and Wenderoth, 2005) an adheres to the principled idea of learning ranked qualia structures. In fact, a ranking of qualia elements is useful as it helps to determine a cut-off point and as a reliability indicator for lexicographers inspecting the qualia structures. In contrast to our previous work, the focus of this paper lies in analyzing different measures for ranking the qualia elements in the automatically acquired qualia structures. We also introduce additional patterns for the agentive role which make use of wildcard operators. Further, we present a gold standard for qualia structures created for the 30 words used in the evaluation of Yamada and Baldwin (Yamada and Baldwin, 2004). The evaluation

[1]The work reported in this paper has been supported by the X-Media project, funded by the European Commission under EC grant number IST-FP6-026978 as well by the SmartWeb project, funded by the German Ministry of Research. Thanks to all our colleagues for helping to evaluate the approach.

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 888–895,
Prague, Czech Republic, June 2007.

presented here is thus much more extensive than our previous one (Cimiano and Wenderoth, 2005), in which only 7 words were used. We present a quantitative evaluation of our approach and a comparison of the different ranking measures with respect to this gold standard. Finally, we also provide an evaluation in which test persons were asked to inspect and rate the learned qualia structures a posteriori. The paper is structured as follows: Section 2 introduces qualia structures for the sake of completeness and describes the specific structures we aim to acquire. Section 3 describes our approach in detail, while Section 4 discusses the ranking measures used. Section 5 then presents the gold standard as well as the qualitative evaluation of our approach. Before concluding, we discuss related work in Section 6.

2 Qualia Structures

In the Generative Lexicon (GL) framework (Pustejovsky, 1991), Pustejovsky reused Aristotle's basic factors (i.e. the material, agentive, formal and final causes) for the description of the meaning of lexical elements. In fact, he introduced so called *qualia structures* by which the meaning of a lexical element is described in terms of four roles: *Constitutive* (describing physical properties of an object, i.e. its weight, material as well as parts and components), *Agentive* (describing factors involved in the bringing about of an object, i.e. its creator or the causal chain leading to its creation), *Formal* (describing properties which distinguish an object within a larger domain, i.e. orientation, magnitude, shape and dimensionality), and *Telic* (describing the purpose or function of an object).

Most of the qualia structures used in (Pustejovsky, 1991) however seem to have a more restricted interpretation. In fact, in most examples the *Constitutive* role seems to describe the parts or components of an object, while the *Agentive* role is typically described by a verb denoting an action which typically brings the object in question into existence. The *Formal* role normally consists in typing information about the object, i.e. its hypernym. In our approach, we aim to acquire qualia structures according to this restricted interpretation.

3 Automatically Acquiring Qualia Structures

Our approach to learning qualia structures from the Web is on the one hand based on the assumption that instances of a certain semantic relation can be acquired by matching certain lexico-syntactic patterns more or less reliably conveying the relation of interest in line with the seminal work of Hearst (Hearst, 1992), who defined patterns conveying hyponym/hypernym relations. However, it is well known that Hearst-style patterns occur rarely, such that matching these patterns on the Web in order to alleviate the problem of data sparseness seems a promising solution. In fact, in our case we are not only looking for the hypernym relation (comparable to the *Formal*-role) but for similar patterns conveying a *Constitutive*, *Telic* or *Agentive* relation. Our approach consists of 5 phases; for each *qualia term* (the word we want to find the qualia structure for) we:

1. generate for each qualia role a set of so called *clues*, i.e. search engine queries indicating the relation of interest,
2. download the snippets (abstracts) of the 50 first web search engine results matching the generated clues,
3. part-of-speech-tag the downloaded snippets,
4. match patterns in the form of regular expressions conveying the qualia role of interest, and
5. weight and rank the returned qualia elements according to some measure.

The patterns in our pattern library are actually tuples (p, c) where p is a regular expression defined over part-of-speech tags and c a function $c : string \rightarrow string$ called the *clue*. Given a nominal n and a clue c, the query $c(n)$ is sent to the web search engine and the abstracts of the first m documents matching this query are downloaded. Then the snippets are processed to find matches of the pattern p. For example, given the clue $f(x) =$ "$such\ as\ p(x)$" and the qualia term *computer* we would download m abstracts matching the query f(computer), i.e. "such as computers". Hereby $p(x)$ is a function returning the plural form of x. We implemented this function as a lookup in a lexicon in which plural nouns are mapped to their base form. With the use of such clues, we thus download a number of snippets returned by the web search engine in which a corresponding regular expression will probably be matched, thus restricting the linguistic analysis to a few promising pages. The downloaded abstracts are then part-of-speech tagged using QTag (Tufis and Mason, 1998). Then we match the corresponding pattern p in the downloaded snippets thus yielding candidate qualia elements as output. The qualia elements are then ranked according to some measure (compare Section 4), resulting in what we call *Ranked Qualia Structures* (RQSs). The clues and patterns used for the different roles can be found in Tables 1 - 4. In the specification of the clues, the function a(x) returns the appropriate indefinite article – *'a'* or *'an'* – or no article at all for the noun x. The use of an indefinite article or no article at all accounts for the distinction between countable nouns (e.g. such as knife) and mass nouns (e.g. water). The choice between using the articles *'a'*, *'an'* or no article at all is determined by issuing appropriate queries to the web search engine and choosing the article leading to the highest number of results. The corresponding patterns are then matched in the 50 snippets returned by the search engine for each clue, thus leading to up to 50 potential qualia elements per clue and pattern[2]. The patterns are actually defined over part-of-speech tags. We indicate POS-tags in square brackets. However, for the sake of simplicity, we largely omit the POS-tags for the lexical elements in the patterns described in Tables 1 - 4. Note that we use traditional regular expression operators such as $*$ (sequence), $+$ (sequence with at least one element) | (alternative) and ? (option). In general, we define a noun phrase (NP) by the following regular expression: NP:=[DT]? ([JJ])+? $\underline{\text{[NN(S?)])+}}$[3], where the head is the underlined expression, which is lemmatized and considered as a candidate qualia element. For all the patterns described in this section, the underlined part corresponds to the extracted qualia element. In the patterns for the formal role (compare Table 1), NP_{QT} is a noun phrase with the qualia term as head, whereas NP_F is a noun phrase with the potential qualia element as head. For the constitutive role patterns, we use a noun phrase variant NP' defined by the regular expression NP':= (NP of[IN])? $\underline{\text{NP (, NP)* ((,)? (and|or) NP)?}}$, which allows to extract enumerations of constituents (compare Table 2). It is important to mention that in the case of expressions such as *"a car comprises a fixed number of basic components"*, *"data mining comprises a range of data analysis techniques"*, *"books consist of a series of dots"*, or *"a conversation is made up of a series of observable interpersonal exchanges"*, only the NP after the preposition 'of' is taken into account as qualia element. The *Telic* Role is in principle acquired in the same way as the *Formal* and *Constitutive* roles with the exception that the qualia element is not only the head of a noun phrase, but also a verb or a verb followed by a noun phrase. Table 3 gives the corresponding clues and patterns. In particular, the returned candidate qualia elements are the lemmatized underlined expressions in PURP:=$\underline{\text{[VB] NP}}$ | $\underline{\text{NP}}$ | be$\underline{\text{[VBD]}}$. Finally, concerning the clues and patterns for the agentive role shown in Table 4, it is interesting to emphasize the usage of the adjectives 'new' and 'complete'. These adjectives are used in the patterns to increase the expectation for the occurrence of a creation verb. According to our experiments, these patterns are indeed more reliable in finding appropriate qualia elements than the alternative version without the adjectives 'new' and 'complete'. Note that in all patterns, the participle (VBD) is always reduced to base form (VB) via a lexicon lookup. In general, the patterns have been crafted by hand, testing and refining them in an iterative process, paying attention to maximize their coverage but also accuracy. In the future, we plan to exploit an approach to automatically learn the patterns.

Clue	Pattern
Singular	
"a(x) x is a kind of "	NP_{QT} is a kind of NP_F
"a(x) x is"	NP_{QT} is a kind of NP_F
"a(x) x and other"	NP_{QT} (,)? and other NP_F
"a(x) x or other"	NP_{QT} (,)? or other NP_F
Plural	
"such as p(x)"	NP_F such as NP_{QT}
"p(x) and other"	NP_{QT} (,)? and other NP_F
"p(x) or other"	NP_{QT} (,)? or other NP_F
"especially p(x)"	NP_F (,)? especially NP_{QT}
"including p(x)"	NP_F (,)? including NP_{QT}

Table 1: Clues and Patterns for the *Formal* role

[2]For the constitutive role these can be even more due to the fact that we consider enumerations.

[3]Though Qtag uses another part-of-speech tagset, we rely on the well-known Penn Treebank tagset for presentation purposes.

Clue	Pattern
	Singular
"a(x) x is made up of "	NP_{QT} is made up of NP'_C
"a(x) x is made of"	NP_{QT} is made of NP'_C
"a(x) x comprises"	NP_{QT} comprises (of)? NP'_C
"a(x) x consists of"	NP_{QT} consists of NP'_C
	Plural
"p(x) are made up of "	NP_{QT} is made up of NP'_C
"p(x) are made of"	NP_{QT} are made of NP'_C
"p(x) comprise"	NP_{QT} comprise (of)? NP'_C
"p(x) consist of"	NP_{QT} consist of NP'_C

Table 2: Clues and Patterns for the *Constitutive* Role

Clue	Pattern
	Singular
"purpose of a(x) x is"	purpose of (a|an) x is (to)? PURP
"a(x) is used to"	(a|an) x is used to PURP
	Plural
"purpose of p(x) is"	purpose of p(x) is (to)? PURP
"p(x) are used to"	p(x) are used to PURP

Table 3: Clues and Patterns for the *Telic* Role

4 Ranking Measures

In order to rank the different qualia elements of a given qualia structure, we rely on a certain ranking measure. In our experiments, we analyze four different ranking measures. On the one hand, we explore measures which use the Web to calculate the correlation strength between a qualia term and its qualia elements. These measures are Web-based versions of the Jaccard coefficient (Web-Jac), the Pointwise Mutual Information (Web-PMI) and the conditional probability (Web-P). We also present a version of the conditional probability which does not use the Web but merely relies on the counts of each qualia element as produced by the lexico-syntactic patterns (P-measure). We describe these measures in the following.

4.1 Web-based Jaccard Measure (Web-Jac)

Our web-based Jaccard (Web-Jac) measure relies on the web search engine to calculate the number of documents in which x and y co-occur close to each other, divided by the number of documents each one occurs, i.e.

$$\textit{Web-Jac}(x,y) := \frac{Hits(x * y)}{Hits(x) + Hits(y) - Hits(x\ AND\ y)}$$

So here we are relying on the wildcard operator '*' provided by the Google search engine API[4]. Though the specific function of the '*' operator as implemented by Google is actually unknown, the behavior is similar to the formerly available Altavista NEAR operator[5].

Clue	Pattern
	Singular
"to * a(x) new x"	to [RB]? [VB] a? new x
"to * a(x) complete x"	to [RB]? [VB] a? complete x
"a(x) new has been *"	a? new x has been [VBD]
"a(x) complete x has been *"	a? complete has been [VBD]
	Plural
"to * new p(x)"	to [RB]? [VB] new p(x)
"to * complete p(x)"	to [RB]? [VB] complete p(x)

Table 4: Clues and Patterns for the *Agentive* Role

4.2 Web-based Pointwise Mutual Information (Web-PMI)

In line with Magnini et al. (Magnini et al., 2001), we define a PMI-based measure as follows:

$$Web-PMI(x,y) := log_2 \frac{Hits(x\ AND\ y)\ \mathit{MaxPages}}{Hits(y)\ Hits(y)}$$

where maxPages is an approximation for the maximum number of English web pages[6].

4.3 Web-based Conditional Probability (Web-P)

The conditional probability $P(x|y)$ is essentially the probability that x is true given that y is true, i.e.

$$\textit{Web-P}(x,y) := P(x|y) = \frac{P(x,y)}{P(y)} = \frac{Hits(x\ NEAR\ y)}{Hits(y)}$$

whereby $Hits(x\ NEAR\ y)$ is calculated as mentioned above using the '*' operator. In contrast to the measures described above, this one is asymmetric so that order indeed matters. Given a qualia term qt as well as a qualia element qe we actually calculate Web-P(qe,qt) for a specific qualia role.

4.4 Conditional Probability (P)

The non web-based conditional probability essentially differs from the Web-based conditional probability in that we only rely on the qualia elements

[4] In fact, for the experiments described in this paper we rely on the Google API.

[5] Initial experiments indeed showed that counting pages in which the two terms occur near each other in contrast to counting pages in which they merely co-occur improved the results of the Jaccard measure by about 15%.

[6] We determine this number experimentally as the number of web pages containing the words 'the' and 'and'.

matched. On the basis of these, we then calculate the probability of a certain qualia element given a certain role on the basis of its frequency of appearance with respect to the total number of qualia elements derived for this role, i.e. we simply calculate $P(qe|qr, qt)$ on the basis of the derived occurrences, where qt is a given qualia term, qr is the specific qualia role and qe is a qualia element.

5 Evaluation

In this section, we first of all describe our evaluation measures. Then we describe the creation of the gold standard. Further, we present the results of the comparison of the different ranking measures with respect to the gold standard. Finally, we present an *'a posteriori'* evaluation showing that the qualia structures learned are indeed reasonable.

5.1 Evaluation Measures

As our focus is to compare the different measures described above, we need to evaluate their corresponding rankings of the qualia elements for each qualia structure. This is a similar case to evaluating the ranking of documents within information retrieval systems. In fact, as done in standard information retrieval research, our aim is to determine for each ranking the precision/recall trade-off when considering more or less of the items starting from the top of the ranked list. Thus, we evaluate our approach calculating precision at standard recall levels as typically done in information retrieval research (compare (Baeza-Yates and Ribeiro-Neto, 1999)). Hereby the 11 standard recall levels are 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%. Further, precision at these standard recall levels is calculated by interpolating recall as follows: $P(r_j) = max_{r_j \leq r \leq r_{j+1}} P(r)$, where, $j \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. This way we can compare the precision over standard recall figures for the different rankings, thus observing which measure leads to the better precision/recall trade-off.

In addition, in order to provide one single value to compare, we also calculate the F-Measure corresponding to the best precision/recall trade-off for each ranking measure. This F-Measure thus corresponds to the best cut-off point we can find for the items in the ranked list. In fact, we use the well-known F_1 measure corresponding to the harmonic mean between recall and precision:

$$F_1 := max_j \frac{2\ P(r_j)\ r_j}{P(r_j) + r_j}$$

As a baseline, we compare our results to a naive strategy without any ranking, i.e. we calculate the F-Measure for all the items in the (unranked) list of qualia elements. Consequently, for the rankings to be useful, they need to yield higher F-Measures than this naive baseline.

5.2 Gold Standard

The gold standard was created for the 30 words used already in the experiments described in (Yamada and Baldwin, 2004): *accounting, beef, book, car, cash, clinic, complexity, counter, county, delegation, door, estimate, executive, food, gaze, imagination, investigation, juice, knife, letter, maturity, novel, phone, prisoner, profession, review, register, speech, sunshine, table.* These words were distributed more or less uniformly between 30 participants of our experiment, making sure that three qualia structures for each word were created by three different subjects. The participants, who were all non-linguistics, received a short instruction in the form of a short presentation explaining what qualia structures are, the aims of the experiment as well as their specific task. They were also shown some examples for qualia structures for words not considered in our experiments. Further, they were asked to provide between 5 and 10 qualia elements for each qualia role. The participants completed the test via e-mail. As a first interesting observation, it is worth mentioning that the participants only delivered 3-5 qualia elements on average depending on the role in question. This shows already that participants had trouble in finding different qualia elements for a given qualia role. We calculate the agreement for the task of specifying qualia structures for a particular term and role as the averaged pairwise agreement between the qualia elements delivered by the three subjects, henceforth S_1, S_2 and S_3 as:

$$Agr := \frac{\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} + \frac{|S_1 \cap S_3|}{|S_1 \cup S_3|} + \frac{|S_2 \cup S_3|}{|S_2 \cap S_3|}}{3}$$

Averaging over all the roles and words, we get an average agreement of 11.8%, i.e. our human test subjects coincide in slightly more than every 10th qualia element. This is certainly a very low agreement and certainly hints at the fact that the task considered is certainly difficult. The agreement was lowest (7.29%) for the telic role.

A further interesting observation is that the lowest agreement is yielded for more abstract words, while the agreement for very concrete words is reasonable. For example, the five words with the highest agreement are indeed concrete things: *knife* (31%), *cash* (29%), *juice* (21%), *car* (20%) and *door* (19%). The words with an agreement below 5% are *gaze*, *prisoner*, *accounting*, *maturity*, *complexity* and *delegation*. In particular, our test subjects had substantial difficulties in finding the purpose of such abstract words. In fact, the agreement on the telic role is below 5% for more than half of the words.

In general, this shows that any automatic approach towards learning qualia structures faces severe limits. For sure, we can not expect the results of an automatic evaluation to be very high. For example, for the telic role of *'clinic'*, one test subject specified the qualia element *'cure'*, while another one specified *'cure disease'*, thus leading to a disagreement in spite of the obvious agreement at the semantic level. In this line, the average agreement reported above has in fact to be regarded as a lower bound for the actual agreement. Of course, our approach to calculating agreement is too strict, but in absence of a clear and computable definition of semantic agreement, it will suffice for the purposes of this paper.

5.3 Gold Standard Evaluation

We ran experiments calculating the qualia structure for each of the 30 words, ranking the resulting qualia elements for each qualia structure using the different measures described in Section 4.

Figure 1 shows the best F-Measure corresponding to a cut-off leading to an optimal precision/recall trade-off. We see that the P-measure performs best, while the Web-P measure and the Web-Jac measure follow at about 0.05 and 0.2 points distance. The PMI-based measure indeed leads to the worst F-Measure values.

Indeed, the P-measure delivered the best results for the formal and agentive roles, while for the constitutive and telic roles the Web-Jac measure performed best. The reason why PMI performs so badly is the fact that it favors too specific results which are unlikely to occur as such in the gold standard. For example, while the conditional probability ranks highest: *explore, help illustrate, illustrate* and *enrich* for the telic role of *novel*, the PMI-based measure ranks highest: *explore great themes*, *illustrate theological points*, *convey truth*, *teach reading skills* and *illustrate concepts*. A series of significance tests (paired Student's t-test at an α-level of 0.05) showed that the three best performing measures (P, Web-P and Web-Jaccard) show no real difference among them, while all three show significant difference to the Web-PMI measure. A second series of significance tests (again paired Student's t-test at an α-level of 0.05) showed that all ranking measures indeed significantly outperform the baseline, which shows that our rankings are indeed reasonable. Interestingly, there seems to be an interesting positive correlation between the F-Measure and the human agreement. For example, for the best performing ranking measure, i.e. the P-measure, we get an average F-Measure of 21% for words with an agreement over 5%, while we get an F-Measure of 9% for words with an agreement below 5%. The reason here probably is that those words and qualia elements for which people are more confident also have a higher frequency of appearance on the Web.

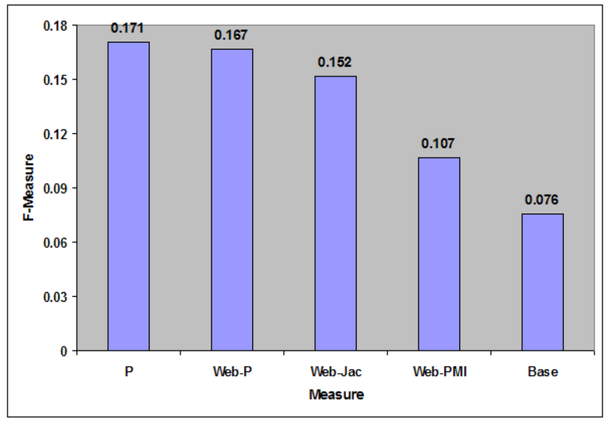

Figure 1: Average F_1 measure for the different ranking measures

5.4 A posteriori Evaluation

In order to check whether the automatically learned qualia structures are reasonable from an intuitive point of view, we also performed an a posteriori evaluation in the lines of (Cimiano and Wenderoth, 2005). In this experiment, we presented the top 10 ranked qualia elements for each qualia role for 10 randomly selected words to the different test persons. Here we only used the P-measure for ranking as it performed best in our previous evaluation with regard to the gold standard. In order to verify that our sample is not biased, we checked that the F-Measure yielded by our 10 randomly selected words (17.7%) does not differ substantially from the overall average F-Measure (17.1%) to be sure that we have chosen words from all F-Measure ranges. In particular, we asked different test subjects which also participated in the creation of the gold standard to rate the qualia elements with respect to their appropriateness for the qualia term using a scale from 0 to 3, whereby 0 means 'wrong', 1 'not totally wrong', 2 'acceptable' and 3 'totally correct'. The participants confirmed that it was easier to validate existing qualia structures than to create them from scratch, which already corroborates the usefulness of our automatic approach. The qualia structure for each of the 10 randomly selected words was validated independently by three test persons. In fact, in what follows we always report results averaged for three test subjects. Figure 2 shows the average values for different roles. We observe that the constitutive role yields the best results, followed by the formal, telic and agentive roles (in this order). In general, all results are above 2, which shows that the qualia structures produced are indeed acceptable. Though we do not present these results in more detail due to space limitations, it is also interesting to mention that the F-Measure calculated with respect to the gold standard was in general highly correlated with the values assigned by the human test subjects in this *a posteriori* validation.

6 Related Work

Instead of matching Hearst-style patterns (Hearst, 1992) in a large text collection, some researchers have recently turned to the Web to match these patterns such as in (Markert et al., 2003) or (Etzioni et al., 2005). Our approach goes further in that it not only learns typing, superconcept or instance-of relations, but also *Constitutive*, *Telic* and *Agentive* relations.

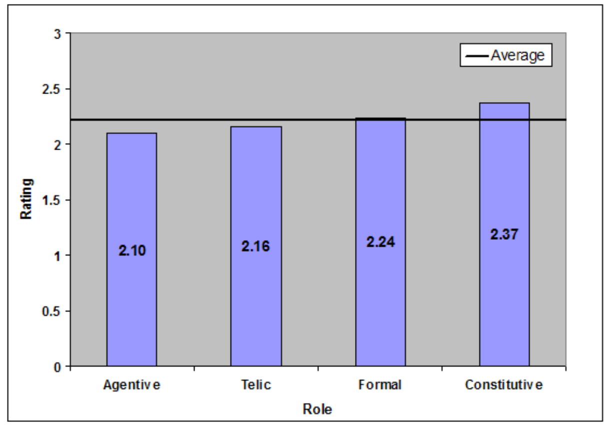

Figure 2: Average ratings for each qualia role

There also exist approaches specifically aiming at learning qualia elements from corpora based on machine learning techniques. Claveau et al. (Claveau et al., 2003) for example use Inductive Logic Programming to learn if a given verb is a qualia element or not. However, their approach does no go as far as learning the complete qualia structure for a lexical element as in our approach. Further, in their approach they do not distinguish between different qualia roles and restrict themselves to verbs as potential fillers of qualia roles.

Yamada and Baldwin (Yamada and Baldwin, 2004) present an approach to learning *Telic* and *Agentive* relations from corpora analyzing two different approaches: one relying on matching certain lexico-syntactic patterns as in the work presented here, but also a second approach consisting in training a maximum entropy model classifier. The patterns used by (Yamada and Baldwin, 2004) differ substantially from the ones used in this paper, which is mainly due to the fact that search engines do not provide support for regular expressions and thus instantiating a pattern as 'V[+ing] Noun' is impossible in our approach as the verbs are unknown a priori.

Poesio and Almuhareb (Poesio and Almuhareb, 2005) present a machine learning based approach to classifying attributes into the six categories: *quality*, *part*, *related-object*, *activity*, *related-agent* and *non-attribute*.

7 Conclusion

We have presented an approach to automatically learning qualia structures from the Web. Such an approach is especially interesting either for lexicographers aiming at constructing lexicons, but even more for natural language processing systems relying on deep lexical knowledge as represented by qualia structures. In particular, we have focused on learning ranked qualia structures which allow to find an ideal cut-off point to increase the precision/recall trade-off of the learned structures. We have abstracted from the issue of finding the appropriate cut-off, leaving this for future work. In particular, we have evaluated different ranking measures for this purpose, showing that all of the analyzed measures (Web-P, Web-Jaccard, Web-PMI and the conditional probability) significantly outperformed a baseline using no ranking measure. Overall, the plain conditional probability P (not calculated over the Web) as well as the conditional probability calculated over the Web (Web-P) delivered the best results, while the PMI-based ranking measure yielded the worst results. In general, our main aim has been to show that, though the task of automatically learning qualia structures is indeed very difficult as shown by our low human agreement, reasonable structures can indeed be learned with a pattern-based approach as presented in this paper. Further work will aim at inducing the patterns automatically given some seed examples, but also at using the automatically learned structures within NLP applications. The created qualia structure gold standard is available for the community[7].

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING/ACL'98*, pages 86–90.

J. Bos, P. Buitelaar, and M. Mineur. 1995. Bridging as coercive accomodation. In *Working Notes of the Edinburgh Conference on Computational Logic and Natural Language Processing (CLNLP-95)*.

P. Cimiano and J. Wenderoth. 2005. Learning qualia structures from the web. In *Proceedings of the ACL Workshop on Deep Lexical Acquisition*, pages 28–37.

V. Claveau, P. Sebillot, C. Fabre, and P. Bouillon. 2003. Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming. *Journal of Machine Learning Research*, (4):493–525.

O. Etzioni, M. Cafarella, D. Downey, A-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

C. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING'92*, pages 539–545.

M. Johnston and F. Busa. 1996. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX workshop on breadth and depth of semantic lexicons*.

A. Kilgariff and G. Grefenstette, editors. 2003. *Special Issue on the Web as Corpus of the Journal of Computational Linguistics*, volume 29(3). MIT Press.

B. Magnini, M. Negri, R. Prevete, and H. Tanev. 2001. Is it the right answer?: exploiting web redundancy for answer validation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 425–432.

K. Markert, N. Modjeska, and M. Nissim. 2003. Using the web for nominal anaphora resolution. In *Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*.

M. Poesio and A. Almuhareb. 2005. Identifying concept attributes using a classifier. In *Proceedings of the ACL Workshop on Deep Lexical Acquisition*, pages 18–27.

J. Pustejovsky, P. Anick, and S. Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Lingustics, Special Issue on Using Large Corpora II*, 19(2):331–358.

J. Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4):209–441.

D. Tufis and O. Mason. 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. In *Proceedings of LREC*, pages 589–96.

E.M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69.

I. Yamada and T. Baldwin. 2004. Automatic discovery of telic and agentive roles from corpus data. In *Proceedings of the the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

[7]See http://www.cimiano.de/qualia.