

Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies

Andreas Wagner

SFB 441

Universität Tübingen

wagner@sfs.uni-tuebingen.de

Abstract

This paper deals with methods for estimating frequencies of concepts in wordnets from corpus data. In particular, it addresses issues which multiple inheritance structures in wordnets raise regarding this task. One of the discussed approaches (tree cut) is problematic in this respect, because it requires a pure tree hierarchy. Applying this approach to a wordnet requires that its DAG structure is transformed into a tree. I propose a mathematically sound method for that purpose and compare this method to a commonly used ad-hoc strategy. This strategy leads to biases in the estimated frequencies which are avoided by the approach proposed here. Experiments with GermaNet demonstrate that these biases have significant impacts.

1 Introduction

Wordnets, i.e. lexical-semantic hierarchies in the style of WordNet (cf. (Fellbaum, 1998)), have commonly been employed in NLP applications which involve quantitative methods. In particular, within the paradigm of statistical corpus linguistics, approaches have been proposed which combine the quantitative evidence provided by word frequencies obtained from a corpus with the symbolic knowledge provided by a wordnet. To establish this combination, the frequencies of

words in the corpus are propagated to the respective concepts that subsume these words. In this way, concept frequencies are estimated from word frequencies. For example, the frequency of the word “person” in the corpus plays a role for the frequency estimates for the concepts <person>, <life_form>, and <entity> in the semantic hierarchy. Concept frequencies, in turn, are used to estimate concept probabilities, which then can be employed for the NLP task in question.

A fundamental issue in this context is *how* concept frequencies can be adequately estimated from word frequencies. This paper is concerned with this issue. In principle, there are several possible ways to achieve that goal. In section 2, I will sketch three basic methods and discuss suitability conditions for their application by considering approaches to a particular NLP task. It turns out that different acquisition approaches—even if they serve the same task—demand different methods of estimating concept frequencies.

The rest of the paper focuses on a general incompatibility that arises if one of the methods described in section 2 is applied to a wordnet. This method requires that the concept hierarchy has a pure tree structure. However, a wordnet generally has the structure of a DAG, i.e. a concept may have more than one parent (immediate hyperonym). To overcome this conflict, a simple ad-hoc strategy to (virtually) convert the DAG structure into a tree structure has been largely used. In section 3, I will point out that this strategy introduces undesirable biases into the frequency estimations. Treating multiple inheritance in an ad-hoc manner

has been justified (if at all) by the fact that multiple inheritance (multiple parents) in WordNet is rare: Less than 1% of the noun concepts in WordNet have more than one parent, most of which are very specific, i.e. located at low levels of the hierarchy (cf. (McCarthy, 2001)). However, for other wordnets, the situation is different. For example, for GermaNet (cf. (Hamp and Feldweg, 1997), (Kunze and Wagner, 2001)), cross-classification of concepts has been a major design principle, and thus multiple inheritance is common; 11.5% of the GermaNet concepts have more than one parent. Hence, when applying a frequency estimation method which requires a tree-shaped hierarchy to a hierarchy like GermaNet, a principled solution to that conflict is highly desirable. Therefore, I propose a more sophisticated method for propagating word frequency counts to concepts. This method converts a wordnet DAG structure into a tree structure, but avoids the drawbacks mentioned above.

Finally, in section 4, I report some experiments performed with GermaNet. These experiments show that the biases introduced by the abovementioned ad-hoc strategy have significant impacts.

2 Basic Methods

2.1 An exemplary task

In order to exemplify the use of different ways to estimate concept frequencies, I will discuss their role in a particular task: learning selectional preferences. Selectional preferences are semantic preferences that a predicate (e.g. a verb) exhibits for its arguments. For example, the verb “eat” prefers a subject referring to a human being or animal and an object denoting food. Such preferences can be represented by wordnet concepts. Statistical approaches for acquiring selectional preferences using WordNet retrieve for each concept a preference value which quantifies the degree of preference (or dispreference) of that concept (with regard to a certain argument slot of a certain verb). The computation of such preference values is based on concept probabilities, which are derived from concept frequencies.

In this section, I describe the basic approaches for concept frequency estimation which have been proposed in the literature that deals with learn-

ing selectional preferences by combining statistical corpus analysis and WordNet. Furthermore, I sketch how these frequency counts are employed for preference acquisition. It turns out that different ways to choose the concepts that should *represent* the selectional preferences of a verb (e.g. <food> for the object of “eat”) require different frequency estimation strategies.

The training data that are used by the approaches discussed here are extracted from a parsed corpus. They comprise pairs of the form (v, n) , where v is a verb and n is the head noun of a certain fixed argument type (e.g. the object) of v . From these data, the verb–noun pair frequencies $freq(v, n)$ as well as the marginal frequencies $freq(v)$ and $freq(n)$ (the overall frequencies of v and n in the data) are extracted and employed to estimate noun concept frequencies $freq(ncpt)$ and $freq(v, ncpt)$, respectively, where $ncpt$ is a concept subsuming n . Based on these concept counts, concept probabilities are usually obtained by maximum likelihood estimation:

$$p(ncpt|v) = \frac{freq(v, ncpt)}{freq(v)} \quad (1)$$

$$p(ncpt) = \frac{freq(ncpt)}{N} \quad (2)$$

where N is the size of the training data.

These probabilities are used to obtain the preference value of $ncpt$ (w.r.t. v). There are several ways to quantify selectional preference. Here, I shortly mention the most common ones. The simplest possibility (pursued e.g. in (Li and Abe, 1998)) is to immediately use $p(ncpt|v)$ (the probability that $ncpt$ occurs as complement of v) as preference score. An alternative possibility (proposed in (Abe and Li, 1996)), is to compute the preference value by the ratio

$$\frac{p(ncpt|v)}{p(ncpt)} \quad (3)$$

This quantity measures the probability that $ncpt$ co-occurs with v *relative to* the general probability of $ncpt$ in the data. This definition offers an obvious way to distinguish between preference and dispreference: If the ratio is greater than 1, then v selects $ncpt$ with higher probability than

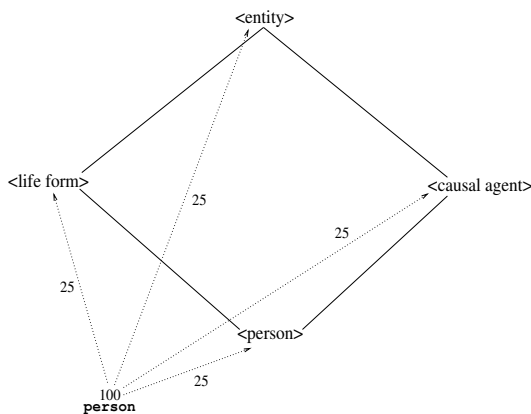


Figure 1: Frequency propagation by the word-to-concept approach

by chance, and thus $ncpt$ is preferred by v . Conversely, a ratio smaller than 1 indicates dispreference.

A third possibility (proposed e.g. in (Resnik, 1998) and (Ribas, 1995)) combines the abovementioned alternatives:

$$p(ncpt|v) \log \frac{p(ncpt|v)}{p(ncpt)} \quad (4)$$

Here, the logarithm of the ratio in (3) (which corresponds to the mutual information between v and $ncpt$) is weighted by $p(ncpt|v)$. Due to the factor $\log \frac{p(ncpt|v)}{p(ncpt)}$, this measure also distinguishes between preferred concepts (preference value > 0) and dispreferred ones (preference value < 0). In addition, the magnitude of the preference score is scaled by the probability that v selects $ncpt$.

2.2 The Word-to-Concept Approach

The method I refer to as word-to-concept approach was proposed by Resnik (cf. (Resnik, 1998)). This method immediately divides the frequency count of a noun n equally among all concepts which subsume n (denoted as $concepts(n)$).

Figure 1 illustrates how the word-to-concept approach works. There are four WordNet concepts that subsume the word “person”: $\langle person \rangle$, $\langle life_form \rangle$, $\langle causal_agent \rangle$, and $\langle entity \rangle$. Thus, each of these four concepts receives $\frac{1}{4}$ of the

frequency of “person” in the corpus ($\frac{100}{4} = 25$ in the example).¹

Formally, the frequency of a concept $ncpt$ is calculated as

$$freq(ncpt) = \sum_{n \in words^+(ncpt)} \frac{freq(n)}{concepts(n)} \quad (5)$$

where $words^+(ncpt)$ is the set of words which are subsumed by $ncpt$, i.e. which are a member either of the synset of $ncpt$ or of the synset of one of its hyponyms. (The joint frequency $freq(v, ncpt)$ of a verb v and a noun concept $ncpt$ is computed analogously; one just replaces $freq(n)$ by $freq(v, n)$ in equation (5).)

The word-to-concept approach yields a probability distribution over *all concepts in the hierarchy*, i.e. the probabilities $p(ncpt)$ of all concepts sum to 1. The same holds for the conditional probabilities $p(ncpt|v)$. This property corresponds to Resnik’s approach to represent the selectional preferences of a verb by *all* WordNet concepts (and their preference values), rather than to retrieve a subset of “representative concepts” for that purpose. Moreover, he uses the distributions $p(ncpt|v)$ and $p(ncpt)$ to quantify the overall *preference strength* of v . The selectional preference strength quantifies how strong the predicate semantically constrains its arguments. For example, “eat” has a greater selectional preference strength for its object than “have”, because “eat” strongly prefers objects denoting food, whereas “have” can select almost any noun as its object. Resnik’s approach of quantifying the overall preference strength is to measure to what extent the probability distribution $p(ncpt|v)$ deviates from the general distribution $p(ncpt)$. The larger the difference between the two distributions, the higher the preference strength. Resnik calculates this difference by the well-known information-theoretic distance measure of *relative entropy*. In fact, (Resnik, 1998) reports a low preference strength for “have” (0.43) and a comparably high preference strength for “eat” (3.51).

¹This is a simplification because it does not take into account that “person” is ambiguous. The example only takes the ‘human’ sense of the word into account. If the data are not lexically disambiguated, which is mostly the case, then the frequency of “person” has to be equally divided among all concepts which subsume any sense of the word.

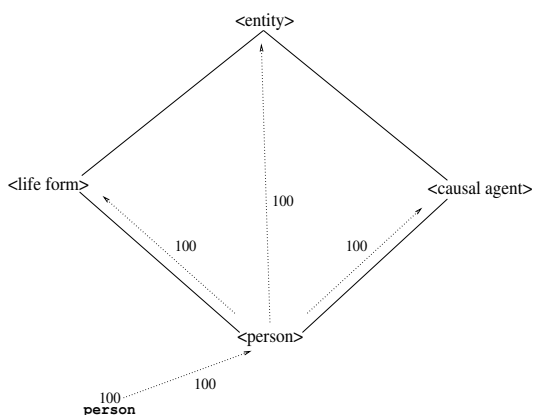


Figure 2: Frequency propagation by the word-to-sense approach

2.3 The Word-to-Sense Approach

While Resnik divides the frequency count of a noun n among all concepts $concepts(n)$ which subsume n , Ribas (cf. (Ribas, 1995)) proposes a different approach: He divides $freq(n)$ among the concepts which represent an immediate sense of n , i.e. those concepts whose synsets contain n (denoted as $senses(n)$). I refer to this strategy as the word-to-sense approach. However, as a noun does not only provide evidence for its senses, but also for the hyperonyms of these senses, the frequency count obtained for a noun sense is completely propagated to each of its ancestors in the hierarchy.

Figure 2 takes up the example in figure 1, this time illustrating the word-to-sense approach. The frequency of “person” (100) is mapped to the synset $\langle person \rangle$, which represents the corresponding word sense.² This count is completely propagated to all concepts that subsume $\langle person \rangle$.

In general, the frequency of a concept is estimated as the sum of the counts of those word senses which the concept subsumes. More formally, let $senses_{ncpt}(n)$ be the set of senses of n which are subsumed by $ncpt$. Then, the frequency

²Again, this simplified example does not take ambiguity into account. If a word is ambiguous (in fact, “person” is) and the data are not disambiguated, then the count of a word is equally divided among its senses.

of a concept is estimated by the equation

$$freq(ncpt) = \sum_{n \in words^+(ncpt)} |senses_{ncpt}(n)| \frac{freq(n)}{|senses(n)|} \quad (6)$$

The word-to-sense approach views the WordNet hierarchy as an inventory of concepts with implication relations among each other. A hyponymy/hyperonymy relation between two concepts indicates that one concept (the hyponym) implies the other (the hypernym). This means that a concept inherits all the probability mass of its hyponyms. In particular, since the root of the hierarchy is implied by all concepts, its probability is 1. In contrast, the word-to-concept approach views the WordNet hierarchy as a pool of concepts which represent a smaller or larger set of nouns. In this model, hyponymy/hyperonymy relations between concepts indicate a common (sub)set of nouns providing evidence for these concepts. This model is required for quantities which are based on probability distributions over the whole inventory of concepts, like Resnik’s overall preference strength. A consequence of this model which might be somewhat counterintuitive is that the probability of the root concept is below 1. This is because probability mass is not completely inherited by, but equally divided among hyperonyms.

As noted above, Ribas quantifies selectional preference according to formula (4). In contrast to Resnik, he does not keep all noun concepts, but extracts a “representative set” of concepts to model the preferential behaviour of a verb. To induce this set, he uses a greedy approach which can be sketched as follows: Initially, consider all noun concepts as “candidates” for inclusion into the representative set. Among them, select that concept $ncpt$ which has the highest preference value and insert it into the target set. After that, remove $ncpt$ and all its hyponyms and hyperonyms from the set of candidates. (This is done to avoid redundancy.) Repeat these steps until the candidate set is empty. In this way, Ribas yields a non-redundant set of highly preferred concepts. For example, (Ribas, 1995) reports that this approach acquired (among others) the following concepts for the subject of “present”: $\langle causal_agent \rangle$ (4.15),

<organization> (0.45), <administrative_district> (0.26), and <life_form> (0.14).

Ribas' simple heuristic for retrieving a representative set of concepts does not depend on a particular approach for estimating concept frequencies. All methods discussed in this paper are compatible with it.

2.4 The Tree Cut Approach

The tree cut approach is a more sophisticated way of retrieving a collection of “representative” concepts from a semantic hierarchy. It was developed by Li and Abe (cf. (Abe and Li, 1996), (Li and Abe, 1998)) for the task of acquiring selectional preferences. Li and Abe represent the selectional preferences of a verb by a *tree cut model*. Such a model provides a horizontal cut through the noun hierarchy so that the concepts which are located along this cut form a partition of the noun senses covered by the hierarchy. A tree cut model consists of the concepts specified by a cut and the preference values for these concepts. Figure 3 shows a portion of the WordNet hierarchy—with preference values attached to the individual concepts, computed according to formula (3)—and two of the possible cuts across the hierarchy (indicated by a solid and a dashed line, respectively). The difference between the corresponding models is that one model contains the concept <animal>, whereas the other model contains the more specific concepts <bird>, <insectivore>, and <primate>. This is an artificial example intended to illustrate plausible preference values and tree cut models for the subject of “fly”.

The tree cut approach aims at finding a cut at the appropriate level of generalisation. In this respect, the cut indicated by the solid line in figure 3 is more appropriate than the more general cut indicated by the dashed line, because the latter one does not capture the fact that some kinds of animals (birds, insects) normally fly, while others do not. The cut at the adequate abstraction level is selected by the *Minimum Description Length (MDL) Principle*. I will not go into details concerning this information-theoretic principle; cf. (Li and Abe, 1998) and (Abe and Li, 1996) for its motivation and application for the given task. In our context, it is important to note that the MDL approach re-

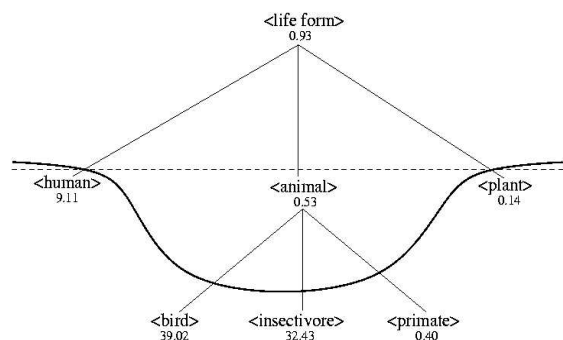


Figure 3: Two possible tree cut models for the subject of “fly”

quires that every possible tree cut model exactly captures the probability mass that represents the whole training data. In other words, the sum of the frequency counts of the concepts on the cut has to correspond to the size of the data.³

To ensure this requirement, the frequency of a noun sense has to be completely propagated to its superconcepts so that the frequency of a concept on the cut (and hence its probability) encompasses the frequencies (probabilities) of all senses it subsumes. Therefore, concept frequencies have to be estimated according to the word-to-sense approach. However, there is a further constraint: It is necessary that each noun sense is subsumed by *one and only one* concept on the cut. Therefore, the structure of the hierarchy must exhibit two properties: Firstly, the noun senses must be modelled by leaf nodes in the hierarchy, while the inner nodes model more abstract concepts. This is required to ensure that all noun senses are below the cut and thus captured by it. Secondly, the hierarchy must be a pure tree, i.e. all concepts (except the root) must have exactly one parent. This is necessary to guarantee that no noun sense is represented

³This requirement follows from the peculiarity that the MDL approach employs tree cut models for *efficiently encoding* the training data, in order to compare the performance of alternative models w.r.t. data compression. This only works properly if all possible tree cut models capture the whole amount of data.

by multiple concepts on the cut.⁴ Obviously, the structure of wordnets deviates from these requirements. Word senses are not only represented by leaves, but by all nodes in the hierarchy. Furthermore, as noted, a wordnet generally exhibits a DAG structure with multiple inheritance.

Thus, to be able to apply the tree cut approach to a wordnet, its structure has to be adapted to meet the two abovementioned properties. To account for the first requirement, a widely used strategy is to create for each inner node an additional node that represents the sense of those words which belong to the synset corresponding to that node. This additional node becomes a hyponym of the original node. In this way, all word senses are captured by leaf nodes. The second requirement is much more complex, since it necessitates a (virtual) transformation of the wordnet DAG structure into a pure tree structure. The core of such a transformation is propagating frequency counts upwards in the hierarchy in a way which “mimics” a tree structure. The next section addresses this issue.

3 Transforming the Wordnet DAG Structure

One crucial part of the virtual transformation of the wordnet structure can be performed as a side effect of processing the hierarchy. If a wordnet is processed top-down (as is done by the tree cut acquisition algorithm developed by Li and Abe), then its DAG structure is “resolved” into a tree structure. Nodes that have multiple parents are processed multiple times, once for each parent. For example, as `<person>` is a hyponym of both `<life_form>` and `<causal_agent>`, this concept (and thus its hyponyms) is processed twice, once as a child of `<life_form>`, and once as a child of `<causal_agent>`. In this way, a “virtual copy” of such a node (and its descendants) is created for each of its parents, and the DAG is “broken into a tree” (cf. figure 4; virtual copies are indicated by a dashed link). Thus, if the task in question involves top-down processing, a tree structure is virtually simulated. Otherwise, the wordnet structure (i.e.

⁴For example, if the cut contained `<life_form>` and `<causal_agent>`, then, assuming the WordNet structure depicted in figures 1 and 2, the senses subsumed by `<person>` would be represented twice.

the database) has to be altered accordingly.

In any case, one has to ensure that the estimated concept frequencies are consistent with that tree structure. As mentioned in section 2.4, the tree cut approach employs the word-to-sense method to obtain concept frequencies, i.e. the frequency of each word sense is propagated to all its ancestors in the hierarchy, and for each concept, the frequencies accumulated at it add up to its count. In fact, there are several possibilities of how to perform this propagation. Following Ribas’ approach explained in section 2.3, the frequency of a concept is the sum of the frequencies of the word senses which are subsumed by that concept (cf. equation (6)). If the hierarchy is a tree structure, then this frequency is equivalent to the sum of the frequencies of the immediate hyponyms (i.e. the children) of the concept:

$$freq(ncpt) = \sum_{ncpt_c \in children(ncpt)} freq(ncpt_c) \quad (7)$$

However, if the hierarchy is a DAG, then equation (7) might yield different values than equation (6). For example, in figure 2, `<entity>` would receive the count of `<life_form>` plus the count of `<causal_agent>`, i.e. the count of `<entity>` would be 200 instead of 100.

A straightforward way to obtain frequency counts consistent with the tree structure is to employ equation (7) instead of equation (6) for frequency estimation. Li and Abe as well as other researchers adopted this solution. Here, the duplication of subtrees is reflected by the corresponding counts. The drawback of this approach is that multiplying certain subtrees corresponds to multiplying that portion of the data which is covered by the concepts in that subtree.

Figure 4 shows an example. Here, as the concept `<person>` is processed twice, all instances in the data denoting a person are counted twice. Thus, the relative proportion of these instances is increased. In particular, the frequency of the top node `<entity>` contains the count of `<person>` twice.

In order to avoid such biases, I propose a different approach for retrieving concept frequencies.

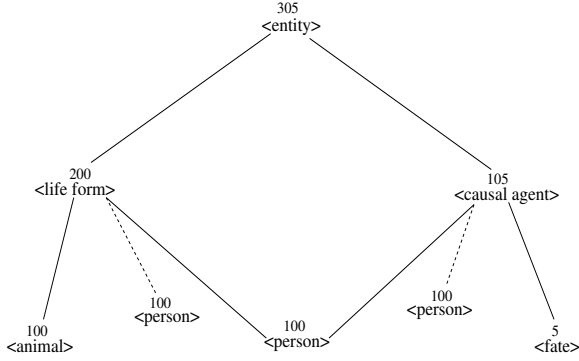


Figure 4: Breaking a DAG into a tree structure

The general idea of this approach is as follows: As in the work of Li and Abe, the count of a concept is directly determined by the counts of its children. This simulates a tree structure. However, a concept does not necessarily inherit the *total* count from each of its children. If a concept has multiple parents, then the count of that concept is divided among its parents. In this way, counts are not duplicated, and thus no bias towards certain parts of the sample is created. The frequency portion that a child concept $ncpt_c$ passes to each of its parents is determined by a probability distribution $p(ncpt_p|ncpt_c)$ where $ncpt_p$ is a parent of $ncpt_c$. Thus, the frequency of a concept is given by

$$freq(ncpt_p) = \sum_{ncpt_c \in children(ncpt_p)} freq(ncpt_c) p(ncpt_p|ncpt_c) \quad (8)$$

The crucial question is how to estimate the distribution $p(ncpt_p|ncpt_c)$ in this equation. I decided to guide this estimation by the frequencies of the parents: The count of a concept is apportioned among its parents according to their respective frequency, relative to the frequencies of the other parents. Formally, for a concept $ncpt_c$, the distribution $p(ncpt_p|ncpt_c)$ is estimated by the ratio of the frequency of $ncpt_p$ and the sum of the

frequencies of all parents of $ncpt_c$:

$$p(ncpt_p|ncpt_c) = \frac{freq(ncpt_p)}{\sum_{ncpt' \in parents(ncpt_c)} freq(ncpt')} \quad (9)$$

In the trivial case in which $ncpt_c$ has only one parent, $p(ncpt_p|ncpt_c)$ is 1, i.e. the complete concept frequency is propagated to that parent.

The equations (8) and (9) depend on each other. The probability of the parent given a child concept in equation (8) is estimated by equation (9), whereas the parent frequencies in equation (9) are obtained by equation (8). Therefore, to make these equations applicable, it is necessary to assume certain initial values. It is quite straightforward to initialise the parent probabilities by assuming uniform distributions:

$$p(ncpt_p|ncpt_c) = \frac{1}{|parents(ncpt_c)|} \quad (10)$$

In this way, the count of a concept is equally apportioned to its parents in the initial iteration. As the parents of a concept have different (additional) children, this iteration yields different counts for them. Thus, in the following iterations, equation (9) will estimate differing probabilities for the parents of a concept. In general, an iteration step changes the counts and probabilities. The approach proposed here can be viewed as an instance of the EM algorithm: equation (8) corresponds to the E-step and equation (9) to the M-step.

For example, in figure 5, the initialisation step equally apports the count of <person> to its two parents; each parent inherits the count $\frac{100}{2} = 50$. Then, in the reestimation step, the <person> count is divided relative to the frequencies of the parents: <life_form> gets $100 \times \frac{150}{150+55} = 73.17$, while <causal_agent> receives $100 \times \frac{55}{150+55} = 26.83$ from <person>. (The counts for <animal> and <fate> are completely propagated to their respective parents.) Note that the count for the top node <entity> does not change. It corresponds to the unbiased total frequency of the data.

In addition, the count of a child concept $ncpt_c$ has to be apportioned among the different (virtual or real) copies of it which emerge from breaking the DAG into a tree. In the tree structure, each copy of $ncpt_c$ has exactly one parent

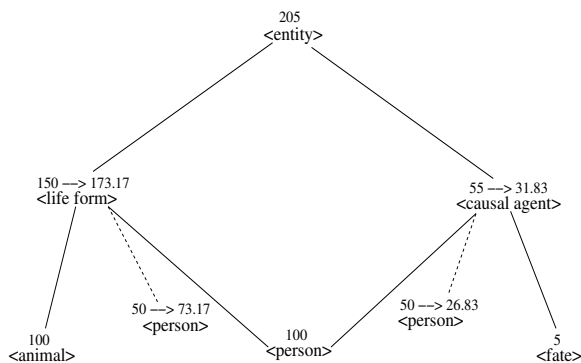


Figure 5: Reestimating frequencies

$ncpt_p$ and receives that portion of the original frequency $freq(ncpt_c)$ that has been propagated to $ncpt_p$, i.e. $freq(ncpt_c)p(ncpt_p|ncpt_c)$. Likewise, the corresponding copies of the descendants of $ncpt_c$ have to be scaled by $p(ncpt_p|ncpt_c)$. Figure 5 illustrates this adjustment for the copies of $\langle person \rangle$.

A possible intuitive access to the general idea that the count of a concept is divided among its parents might be to understand hyperonymy in a more “subjective” manner than usual: Instead of “is a kind of”, a hyperonymy relation could be interpreted as “is perceived / referred to as”. This means that multiple hyperonyms represent different aspects of a concept which might have different salience. For example, a person might be primarily referred to as a life form in some situations (e.g. in an utterance like “How many persons died?”), and as a causal agent in other situations (e.g. in “This person caused the accident.”). The probabilities $p(\langle life_form \rangle | \langle person \rangle)$ and $p(\langle causal_agent \rangle | \langle person \rangle)$, together with the corresponding split of the count of $\langle person \rangle$, model the relative salience of these two aspects w.r.t. $\langle person \rangle$. The way proposed here to estimate these probabilities employs the only empirical quantitative information about the parent concepts that is available: their overall frequency. A parent that has a high frequency (compared to the other parents) gets a high probability, while a parent with a (comparably) low frequency is assigned

a low probability. The count of a parent concept reflects its “global” salience (w.r.t. the training data); the comparison with the counts of the other parents reflects peculiarities of their common child.

More formally, the approach described here can be viewed as performing a soft classification of noun senses. The concepts can be regarded as soft classes of senses, and multiple hyperonymy corresponds to graded membership. For example, all instances of $\langle person \rangle$ are graded members of both classes $\langle life_form \rangle$ and $\langle causal_agent \rangle$. The degree of membership is represented by $p(\langle life_form \rangle | \langle person \rangle)$ and $p(\langle causal_agent \rangle | \langle person \rangle)$, respectively.

4 Experiments

This section describes experiments I carried out to test the effect of employing the two frequency estimation methods sketched in section 3 for acquiring selectional preferences using the tree cut approach. As mentioned, the method using equation (7) (henceforth called ‘Simple’) multiplies frequency counts of noun senses which are covered by duplicated concepts, while the approach using equations (8)–(10) (henceforth called ‘Reestimation’) avoids such a bias. For the experiments, I used GermaNet as semantic hierarchy. As noted in section 1, multiple inheritance is a common structural property in this resource. This suggests that the bias which the Simple approach imposes on the frequency estimates is significant when applied to GermaNet. The experiments described below aim at verifying this hypothesis.

4.1 Setting

The experiments acquired selectional preferences for the object of several verbs. The training data I used were extracted from parsed relative clauses and verb-final clauses originating from a large German newspaper corpus. This parsed corpus was created at the IMS, University of Stuttgart. From these sentences, I extracted verb–noun (object) pairs (666,831 altogether). To avoid the problem of data sparseness, I acquired selectional preferences for those verbs which occur at least 500 times in the training set (261 verbs). For preference acquisition, I used a modified version of the tree cut approach described in (Abe and Li,

1996).⁵ This modification involves an additional parameter that can be varied to influence the generalisation level of the induced cut (cf. (Wagner, 2000) or (Wagner, 2002) for details of this modified approach). With this parameter, I forced the algorithm to select the cut at or close to the highest possible level of abstraction, which comprises the top concepts of GermaNet. This is a conservative proceeding, since differences in tree cuts are much more likely if they tend to be located at low levels in the hierarchy, capturing peculiarities of very specific concepts.

Concerning frequency estimation, I carried out the experiments once using the Simple approach and once using the Reestimation approach (after the initial iteration using equation (10), I performed one reestimation iteration).

4.2 Results

The results show considerable differences between the selectional preferences acquired using the Simple and the Reestimation approach, respectively. First of all, it turned out that Simple yielded significantly higher total frequency counts at the hierarchy root for each verb than Reestimation: The average total count per verb was 1300.35 for Simple vs. 1149.55 for Reestimation. This means that Simple artificially increased the total count of the data by 13%. A more interesting question is to what extent the preferences acquired with the two approaches are different. Comparing the individual concepts which are classified as being preferred (preference value > 1), the difference is considerable. For the whole set of 261 test verbs, Simple acquired 1085 preferred concepts, Reestimation 1087 preferred concepts altogether. Of these, 924 concepts were equal. This amounts to a difference of 15%. At first glance, this does not seem too much. But taking into account that the cuts comprise concepts at a very high generalisation level, the difference is remarkable. Looking at the complete preference profiles acquired for each verb, the picture becomes much more clear-cut. Only for 99 verbs, i.e. 38% of the test verbs, the two methods yielded the same set of preferred concepts.

⁵As mentioned, this approach employs equation (3) to compute preference values.

As an example, table 1 shows the tree cut models acquired for “wissen” (to know). Both models classify the concept $\langle \text{Attribut\#Eigenschaft} \rangle$ (attribute, property) as preferred. The Reestimation cut also models $\langle \text{?kognitives_Objekt} \rangle$ (cognitive object) as preferred concept, which is in accordance with human intuition. The Simple cut does not contain this concept, since it is located one level higher, at $\langle \text{Entität} \rangle$ (entity), which subsumes $\langle \text{?kognitives_Objekt} \rangle$ and $\langle \text{Objekt} \rangle$ (object). However, the Simple model classifies $\langle \text{Zustand} \rangle$ (state) as preferred, which is much less intuitive. The probability distributions $p(\text{ncpt}|\text{wissen})$ of the concepts on the two cuts are rather similar, though some differences (e.g. 0.16 versus 0.20 for $\langle \text{Situation} \rangle$) might matter when employed for a particular application.

Altogether, the experiments verify that the Simple results differ significantly (though not dramatically) from the Reestimation results.

5 Conclusion

In this paper, I discussed different methods for estimating frequencies of concepts in wordnets from corpus data. Based on an example NLP task (selectional preference acquisition), I illustrated that the selection of an appropriate frequency estimation method largely depends on the statistical methods that employ the induced frequencies. In particular, this paper focused on the problems which multiple inheritance in wordnets impose on concept frequency estimation. Two of the discussed methods, word-to-concept and word-to-sense, are suitable for multiple inheritance hierarchies without modification. These approaches rest on the subsumption relation between words and concepts rather than the immediate hyperonymy relation and thus are compatible with DAG structures. However, the tree cut approach requires a concept hierarchy that exhibits a pure tree structure. To apply this approach to a wordnet requires a transformation of the wordnet’s DAG structure. I discussed the most commonly used ad-hoc strategy for this transformation. This strategy leads to biases of the estimated frequency counts, which are evoked just by the multiple inheritance structure. Therefore, I proposed a more sophisticated EM-style strategy which involves the ad-

Reestimation			Simple		
concept	pref. value	prob.	concept	pref. value	prob.
<?kognitives_Objekt>	1.25	0.21	<Entität>	0.69	0.32
<Objekt>	0.42	0.13	<Verhältnis#Relation>	0.61	0.03
<Verhältnis#Relation>	0.94	0.02	<Stelle#Ort#Stätte>	0.42	0.02
<Stelle#Ort#Stätte>	0.36	0.02	<Motiv#Intention>	0.45	0.003
<Motiv#Intention>	0.43	0.004	<Menge>	0.50	0.03
<Menge>	0.66	0.02	<Situation>	0.83	0.20
<Situation>	0.65	0.16	<Besitz>	0.04	0.001
<Besitz>	0.02	0.0006	<Zustand>	1.13	0.02
<Zustand>	0.51	0.009	<Attribut#Eigenschaft>	4.77	0.37
<Attribut#Eigenschaft>	5.10	0.40			

Table 1: Tree cut models for “wissen”

justment and reestimation of frequency counts. Experiments showed that the bias imposed by the ad-hoc approach is significant.

For future work, it will be interesting to test the performance of the different frequency estimation approaches w.r.t. particular NLP tasks. For example, selectional preferences acquired by the two approaches tested in section 4 could be employed for lexical or structural disambiguation. A priori, it is not clear whether the mathematically sound approach which I proposed performs better than the simple ad-hoc approach. This has to be examined empirically. In any case, the issue of concept frequency estimation should not be disregarded.

References

- Naoki Abe and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proc. of 13th Int. Conf. on Machine Learning*.
- Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.
- Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Madrid.
- Claudia Kunze and Andreas Wagner. 2001. Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In Bernhard Schröder, Angelika Storrer, and Ingrid Lemberg, editors, *Probleme und Perspektiven computergestützter Lexikographie*. Niemeyer.
- Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.
- Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex, Brighton.
- Philip Resnik. 1998. WordNet and class-based probabilities. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 239–263. MIT Press, Cambridge, Mass.
- Francesc Ribas. 1995. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy*. PhD thesis, Universitat Politècnica de Catalunya.
- Andreas Wagner. 2000. Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proc. of ECAI-2000 Workshop on Ontology Learning*, pages 37–42, Berlin.
- Andreas Wagner. 2002. Learning thematic role relations for wordnets. In *Proc. of ESSLI 2002 Workshop on Machine Learning Approaches in Computational Linguistics*, pages 99–113, Trento.