Assessing the difficulty of finding people in texts

Constantin Orăsan and Richard Evans

Computational Linguistic Group
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton, WV6 0QB
United Kingdom
{C.Orasan, R.J.Evans}@wlv.ac.uk

Abstract

In this paper several methods for animacy recognition are evaluated. Each method has an increasing complexity over the previous one and involves more resources, and as a result, more computation. When assessing the performace of these methods we consider three factors: the results of an intrinsic evaluation, the results of an extrinsic evaluation, and the complexity of the method. For intrinsic evaluation the accuracy of the overall classification is considered as well as the precision and recall for each type classification. In the extrinsic evaluation, the animacy classifier is used to filter candidates in a pronominal anaphora resolution system. Given the wide variety of texts used, an anaphora resolution system could not be used for this evaluation because its performance depends upon the genre of the text being processed. For this reason, the reduction of the number of candidates, the reduction of the number of antecedents, and the increase in the number of pronouns without any antecedents were recorded and used to differentiate between the systems. Comparison between different systems showed that the best one is the system which uses machine learning, and that the additional information brought by different modules does not lead to an increase in the success of the system due to the errors introduced by them.

1. Introduction

In English, automatic identification of the specific gender of English nouns is a difficult task of arguably limited utility. However, information about the animacy of nouns can be very useful in tasks like anaphora resolution, coreference resolution and parsing. It can also contribute to applications such as question answering, allowing users to obtain answers to "who" questions. For our purposes, a noun phrase is animate if, in its singular form, its referent can be referred to using one of the pronouns *he*, *she*, *him*, *her*, *his*, *himself*, or *herself*.

In the present work, we consider that the animacy of a noun phrase (NP) is derived from the animacy of its head. While the gender of a NP can be specified by adjectival modifiers such as *female* or *male*, animacy is not normally specified in this way. To illustrate, both NPs *the man* and *the dead man* can be referred to using the same animate pronoun. In this way, our treatment of NP animacy mirrors the treatment of grammatical number under the Government and Binding Theory (Chomsky, 1981).

In this paper several methods for animacy recognition are evaluated. Firstly, a simple statistical method based on WordNet (Fellbaum, 1998) is shown to be quite useful for the task, but is unable to classify animate entities with high accuracy. Next, a machine learning method is used in order to improve the results. In the latest stages of development, word sense disambiguation (WSD), lists of names, and a method for named entity recognition (NER) are added to further improve the accuracy of the classification.

The systems are evaluated using intrinsic and extrinsic evaluation methods. For the intrinsic evaluation, each method is considered a classifier and specific measures for classification (e.g. accuracy, precision, and recall) are applied. For extrinsic evaluation, the influence of the systems on anaphora resolution is assessed.

Previous work in anaphora resolution (AR) has shown that levels of performance are related to both the type of text being processed and to the average number of NPs under consideration as a pronoun's antecedent (Evans and Orăsan, 2000). Veins Theory (Cristea et al., 2000) sought to use discourse structure in order to select minimal sets of suitable candidates for a pronoun's antecedent. It is generally recognised that number and gender agreement constraints between pronouns and competing candidates is an effective method for reducing the size of those sets of candidates. In this work, we evaluate the usefulness of animacy agreement constraints as a weaker substitute for the more difficult task of implementing gender agreement constraints on the process of anaphora resolution.

Our implemented system for anaphora resolution, MARS (Mitkov et al., 2002), was designed to resolve pronouns in texts from a technical domain. For every pronoun identified in a text, that system extracts NPs from the preceding part of the sentence in which the pronoun appears as well as two sentences preceding that, as well as the section heading. All NPs failing to agree with the pronoun in number are discarded. The remainder form a set of competing candidates from which MARS should select the antecedent.[1] Intuitively, assuming complete preservation of antecedent NPs in the set of competing candidates, the smaller the set, the better the algorithm's chances of selecting the antecedent from it. Ideally, the size of the set would be reduced further by discarding those NPs that fail to agree with the pronoun in terms of their gender. Since performance in AR is influenced by the domain, and the training corpus used in this paper is derived from quite different domains, we will not present MARS's performance directly in this paper. Instead, we will examine the degree to which the output of the methods benefits a system for anaphora resolution by reducing the

[1] In fact, most systems for pronominal anaphora resolution work in this way.

size of the sets of competing candidates. The degree of preservation of valid antecedents in the resultant sets of candidates will also be taken as a measure of success.

This paper is structured as follows. In Section 2. we describe several methods for identifying the animacy of nouns. In Section 3. the methods are assessed using intrinsic and extrinsic evaluation techniques and a comparison is made of their complexity. Related work is discussed in Section 4. Conclusions are drawn in Section 5.

2. Animacy recognition

In this section several methods previously developed for animacy recognition are briefly presented. Each approach uses increasing amounts of linguistic resources and tries to improve the classification accuracy of previous methods.

2.1. Simple method based on WordNet

WordNet (Fellbaum, 1998) is an electronic lexical resource organized hierarchically by relations between sets of synonyms or near-synonyms called synsets. Each of the four primary classes of content-words, nouns, verbs, adjectives and adverbs are arranged under a small set of top-level hypernyms called unique beginners. In the case of nouns and verbs, the unique beginners are the most general concepts under which the entire set of entries is organized on the basis of hyponymy and entailment.

It was noticed that several hierarchies were of interest with respect to the aim of identifying the animate entities in texts. In the case of nouns, three of the unique beginners are expected to be hypernyms of senses of nouns that refer to animate entities. These are *animal*, reference number (05), *person* (18), and *relation* (24). There are four verb sense hierarchies that allow the inference to be made that their subject NPs should be animate. The unique beginners in these cases are *cognition* (31), *communication* (32), *emotion* (37) and *social* (41). It was clear that a lexical resource arranged in such a hierarchical fashion could be exploited in order to associate the heads of noun phrases with a measure of confidence that the associated NP has either an animate or inanimate referent.

Motivation for the use of WordNet arises from the fact that knowledge as to the animacy of common NPs cannot be readily computed from explicit features of the text. Unlike the situation with proper names and their associated clues such as titles and initial capitalisation, knowledge as to the animacy of common NPs appears to be purely implicit. Recognition of animate entities must, at some point be grounded in world-knowledge - information partially embodied in WordNet.

The animacy of each noun in the text is then decided using the information from WordNet. In each case, WordNet is consulted to examine all the possible senses of a noun. Entries in WordNet are provided with a list of their senses as used in the corpus from which the resource was developed. This means that a method attempting to find the particular sense of a word must do more than simply consult WordNet, it should ideally perform word sense disambiguation (WSD) in order to extract the particular sense of a word in each case. WSD was not incorporated in the simple method. Instead, a count is made of the number of animate senses that is listed for a noun (hyponyms of unique beginners 05, 18, or 24) and the number of inanimate senses (hyponyms of the remaining unique beginners). A ratio is computed and nouns with a ratio greater than a pre-defined threshold are classified as animate. Similarly, in the case of nouns that are the heads of subject NPs, counts are made of the animate and inanimate senses of the verbs they are subjects of. A ratio is then computed and subject nouns of verbs with a ratio higher than some threshold are classified accordingly. Finally, contextual rules (e.g. the presence of NP-internal complementisers and reflexives such as *who* or *herself*) are applied in order to improve the classification.

2.2. Machine learning for animacy recognition

Having noted that not all the hyponyms of particular unique beginners can be uniformly classed as either animate or inanimate, we do consider that more specific senses can be uniformly classified in this way. In (Orăsan and Evans, 2001) we presented a corpus-based method which classifies the synsets from WordNet according to their animacy.

The nouns in a 52 file subset of the SEMCOR corpus (Landes et al., 1998) were manually annotated with animacy information and then used by an automatic system to classify more general (though not top-level) senses[2]. SEMCOR is a resource in which each word has been annotated with the sense from WordNet that it conveys.

The system attempts to classify the senses from WordNet that explicitly appear in the corpus directly, on the basis of the frequency with which they have been manually classified as animate or inanimate. One of our goals was to design a procedure which can also classify senses that are not found in the corpus. To this end, we decided to use a bottom up procedure which starts by classifying the terminal nodes in WordNet and then continues with more general nodes. The terminal nodes are classified using information from the annotated files. When classifying a more general node, we considered that if all the hyponyms of a sense are animate, then the sense itself is animate. This does not always hold due to annotation errors or rare uses of a sense and instead, a statstical measure was used to test the animacy of a more general node. After considering several measures, chi-square seemed the most appropriate.

This classification of senses was useful for determining the animacy of a sense, even those which were not previously found in the corpus, but which are hyponyms of a node that has been classified. However, nouns/verbs whose sense is unknown cannot be classified directly and therefore an additional level of processing was necessary. We used the instance-based learning algorithm available in the TiMBL package (Daelemans et al., 2001) to determine the animacy of nouns. Under this type of learning, all the instances are stored without trying to infer anything from them. At the classification stage, the algorithm compares a previously unseen instance with all the data stored at the training stage. The most frequent class in the k nearest neighbours is assigned as the class to which that instance

[2]Information as to the animacy of subject nouns was used to associate verbs with this information

belongs. In our case the instances used in training and classification contained the lemma of the noun which is to be classified, its number of animate and inanimate senses[3], the number of animate/inanimate senses of the verb in the case of subject nouns, and the ratio of animate singular pronouns to inanimate singular pronouns in the text being processed. The output of the instance-based learning stage is a list of nouns classified according to their animacy.

During the evaluation of the methods presented in Sections 2.1. and 2.2. we noticed many errors when named entities were classified. This is because our systems rely on WordNet which either does not contain many of these named entities, or else the different senses listed do not relate them to named entities. For this reason, we decided to ignore non-sentence-initial capitalised words.

2.3. Word sense disambiguation

It is difficult to disambiguate the possible senses of words in unrestricted texts, but it is not so difficult to identify those senses which are more likely to be used in a text than others. Such information was not considered in the methods presented in 2.2. and 2.1. Instead, in those methods, all the senses were considered to have an equal weight. In order to address this problem, the word sense disambiguation (WSD) method described in (Resnik, 1995) was implemented and used in the classification algorithm. The WSD method computes the weight of each possible sense of each noun by considering the other nouns in a text. These weights were used to compute the number of animate/inanimate senses. Our underlying hypothesis is that the animacy/inanimacy of senses which are more likely to be used in a text should count more than that of improbable senses.

2.4. Proper name lists

The methods presented in Sections 2.1. and 2.2., even with the improvements presented in 2.3., are able to classify common nouns but not proper nouns. Attempts to use WordNet to classify proper nouns failed because many of those items do not appear in WordNet or else, those that do appear, have a substantial number of inanimate senses. To illustrate, the names *Bob* and *Maria* are not assigned any animate senses in WordNet. The systems presented in the previous three sections were enhanced by the inclusion of proper name lists. Whenever a non-sentence-initial capitalised word was encountered, it was searched for in a list of person names and if it was found it was considered an animate entity. At present the list contains more than 180,000 given names and surnames.

One limitation of the use of proper name lists is the fact that they are static. Interestingly, as the size of a gazetteer grows, it becomes more difficult to use effectively because it provides the basis of classification for an increasingly ambiguous collection of words. To illustrate, lists of first names and surnames can include names such as *Bacon*, *Lemon*, *Flint*,[4], *Will* or *Ash*. Month names or astronomical terms become confused with person names, words usually denoting professions are confused with the names of computer equipment, and so on.

2.5. Named entity recognition

The presence of a word in a proper name list does not guarantee that in that context the word really does refer to an animate entity. It is very common for personal names to be used to name companies, months, or places. Therefore a named entity recognition (NER) method should classify a capitalised word as a person or other type of entity.

We are currently developing a system for NER. Here, recognition is a two step process where capitalised words are normalised (Mikheev, 2000) and then normally upper case words are classified as references to persons or non-persons. Although the normalisation task is performed with an accuracy of 99.25%, the animacy classification is only 90.21% accurate. It will be noted that at this early stage, the system performs worse than those developed by participants at MUC-7 (Chinchor, 1998). Preliminary error-analysis indicates that this performance is influenced by the features of the texts used in the evaluation. These texts include sentences in block capital letters and a proportion of the texts are incomplete.

3. Evaluation of the systems

In Section 2. several methods for classifying the animacy of a noun were proposed. These systems can be divided into two classes. The first one includes the systems proposed in Sections 2.1., 2.2., and 2.3. which can classify common nouns, but fail to classify proper nouns correctly. In order to address this problem, these systems are extended to classify proper nouns by incorporating the methods proposed in Sections 2.4. and 2.5. In this section these systems are evaluated using *intrinsic* and *extrinsic* evaluation methods (Sparck-Jones and Galliers, 1996). In addition, the complexity of the systems is considered.

For our evaluation we used two corpora. The first one is a selection of texts from the SEMCOR corpus (Landes et al., 1998) stripped of the sense annotation. These texts were chosen because the nouns contained in them were annotated with their corresponding senses from WordNet and therefore could be used to determine the animacy of synsets. They were used to train our machine learning classifier. In addition, these texts contained a large number of references to animate entities. The second corpus used in our paper is a collection of texts from Amnesty International (AI) which were used in developing the method described in Section 2.1. (Evans and Orăsan, 2000). These texts were chosen because they contained a large number of references to animate entities. The characteristics of these corpora are summarised in Table 1.

The reliability of the evaluation is increased by evaluating the systems on both corpora. For example the thresholds used in the simple method presented in Section 2.1. were determined through direct observation of the data used to develop that method. Therefore, by evaluating the method on the SEMCOR corpus, we could measure its performance on completely unseen data. In addition to this, the texts from SEMCOR are in a completely different genre

[3] As stated earlier, in the cases where the animacy of a sense is not known, it is inferred from its hypernyms.

[4] As used in the names of celebrities *Kevin Bacon*, *Jack Lemon*, or *Keith Flint*.

from AI, allowing us to measure how genre independent the system described in Section 2.1. is. Evaluation raises more serious problems for all of the machine learning systems. As is known, whenever a machine learning method is evaluated, a clear distinction has to be made between training data and testing data.

In the case of the system described in (Orăsan and Evans, 2001), we evaluated the approach using 10-fold cross-validation over the SEMCOR corpus. However, given that we also have the AI corpus, we could evaluate the systems on completely unseen data. In addition, the evaluation of the machine learning methods on the AI corpus was useful in proving that the classification of the sysnsets from WordNet on the basis of information from SEMCOR is also useful when applied to different texts.

3.1. Intrinsic evaluation

Intrinsic evaluation methods measure the accuracy of a system in performing the task which it was designed to carry out. In our case, it is the accuracy of classifying an entity as animate or inanimate. In order to assess the performance of the systems four measures are considered:

$$Accuracy = \frac{Correctly\ classified\ items}{Total\ number\ of\ items} \quad (1)$$

$$Precision = \frac{True\ positives}{True\ positives\ +\ False\ positives} \quad (2)$$

$$Recall = \frac{True\ positives}{True\ positives\ +\ False\ negatives} \quad (3)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

The *accuracy* (1) measures how well a system can correctly classify an entity as animate or inanimate, but it can be misleading because of the large number of inanimate entities in texts. As is clear from Table 1, even though the texts were chosen so as to contain a large number of animate entities, the ratio between the number of animate entities and inanimate entities is approximatively 1 to 7.5 for SEMCOR, and 1 to 4.8 for AI. This means that a method which classifies all entities as inanimate would have an accuracy of 88.21% on SEMCOR and 82.77% on AI. As can be seen in Table 2 these results are not very far from the accuracy obtained by the system described in 2.1. However, as mentioned before, we intend to use the filtering of animate entities for anaphora resolution and therefore, the use of a filter which classifies all the entities as inanimate would be highly detrimental. We have observed a much higher ratio of animate pronouns to inanimate ones in the AI and SEMCOR corpora.

It is clearly important to know how well a system is able to identify animate entities and how well it can identify inanimate entities. In order to measure this, *precision* (2) and *recall* (3) are used. The precision with which a system can identify animate entities is defined as the ratio between the number of entities correctly classified as animate and the total number of entities it classifies as animate (including the wrongly classified ones). A method's recall in classifying animate entities is defined as the ratio between the number of entities correctly classified as animate and the total number of animate entities to be classified. The precision and recall of inanimate classification is defined in a similar manner. The *f-measure* (4) combines the precision and the recall in one value. Several formulas for f-measure were proposed, the one we use gives equal importance to precision and recall.

Table 2 presents the accuracy of the classification, and the recall and precision for classifying the animate and inanimate entities.[5] In addition to the methods presented in Section 2., three baseline methods were introduced. The first one classifies an entity as animate or inanimate on a random basis. The performance of this method is displayed in the column *random baseline* of the table. A second random baseline was introduced because we supposed that the number of gender marked pronouns in a text can be an indication of how likely it is that a particular noun appearing in that text is to be animate or inanimate. In this case the probability of an entity to be animate is proportional to the number of gender marked pronouns in the text and the classification is made on a weighted random basis. A similar rule applies for inanimate entities. This second baseline is referred to in the table as *weighted baseline*. In the same table, in order to facilitate the comparison, we included a method which classifies all the entities as inanimate. This method is referred as the *dummy method*.

As can be noticed, Table 2 does not contain the results of the methods which also incorporate word lists and named entity recognition. Those methods which use lists and named entity recognition try to classify more entities than the other methods. This is because all the methods which use WordNet ignore non-sentence-initial capitalised words, due to our observation that most of them represent named entities and we do not have confidence in information from WordNet when classifying them. They are considered, however, in the extrinsic evaluation.

Table 2 shows that all our methods significantly outperform the baselines used. Close investigation of the table shows that on both corpora the best method is the one which uses machine learning (the one presented in Section 2.2.). It obtains high accuracy when classifying both animate and inanimate entities. As is shown in Section 3.2. similar results are also obtained in the extrinsic evaluation. The machine learning method which uses word sense disambiguation (presented in Section 2.3.), yields very similar levels of performance but as we will argue in Section 3.3. the complexity of any word sense disambiguation method does not make it a good alternative. In terms of accuracy, the simple method performs unexpectedly well given its simplicity. However, it fails to accurately classify animate entities. Moreover, comparison with the dummy method on both files shows that the results of the simple method are not much better, which suggests that the simple method has a bias towards recognition of inanimate entities.

The relatively poor accuracy of the *Simple system* can

[5] The accuracy of the simple method is higher than that reported in our previous papers due to a bug in the evaluation program previously used.

Corpus	No. of words	No. of animate entities	No. of inanimate entities	Total entities
SEMCOR	104612	2321	17380	19701
AI	15767	538	2586	3124

Table 1: The characteristics of the two corpora used

Experiment	Acc	Animacy			Inanimacy		
		Prec	Recall	F-meas	Prec	Recall	F-meas
Random baseline on SEMCOR	50.19%	14.11%	50.49%	22.05%	86.19%	50.14%	63.39%
Random baseline on AI	50.60%	19.37%	52.13%	28.24%	82.11%	50.32%	62.39%
Weighted baseline on SEMCOR	37.62%	8.40%	74.44%	15.09%	88.41%	31.64%	46.60%
Weighted baseline on AI	31.01%	18.07%	76.48%	29.23%	79.27%	20.60%	32.70%
Dummy method on SEMCOR	88.21%	0%	-	-	88.21%	100%	93.73%
Dummy method on AI	82.77%	0%	-	-	82.77%	100%	90.57%
Simple system on SEMCOR	91.42%	88.48%	56.42%	68.90%	91.81%	98.51%	95.04%
Simple system on AI	89.61%	94.79%	52.69%	67.73%	88.93%	99.24%	93.80%
Machine learning system on SEMCOR	97.72%	91.91%	89.99%	90.93%	98.75%	98.57%	98.65%
Machine learning system on AI	98.04%	96.31%	92.19%	94.20%	98.33%	99.26%	98.79%
Machine learning with WSD on SEMCOR	97.51%	89.97%	90.14%	90.05%	98.59%	98.56%	98.57%
Machine learning with WSD on AI	97.85%	95.37%	92.00%	93.65%	98.34%	99.07%	98.70%

Table 2: The results of the classification

be explained by the fact that the unique beginners, which are used as the basis for classification in that method, cover too wide a range of senses for them all to belong to a single animate or inanimate class. They are too general to be used as the basis of accurate classification. Additionally, the rules used to assist classification only provided limited recall in identifying animate entities.

3.2. Extrinsic evaluation

In the previous section we evaluated the performance of the classification method and we saw that even simple methods can achieve high accuracy at the expense of low recall and precision in the classification of animate entities. In computational linguistics, the output of one method is often used as the input for another one, and therefore it is important to know how the results of the first method influence the results of the second. This kind of evaluation is called *extrinsic evaluation*. Given that the identification of animate entities is not very useful on its own, but can be vital for tasks like anaphora resolution, it is necessary to perform extrinsic evaluation too.

In light of this, the influence of animacy recognition on anaphora resolution is evaluated. Given that the performance of different anaphora resolution systems can vary greatly depending on the genre of the text being processed (as reported in (Mitkov, 2002)), we decided not to evaluate the final success rate of an anaphora resolution system. Instead, we computed statistics about the numbers of competing candidates for pronouns given that, in line with intuition and as demonstrated in (Evans and Orăsan, 2000) this can have a big influence on the overall performance of an anaphora resolution system. In addition to the number of candidates eliminated by the filter we recorded the number of antecedents wrongly eliminated and the increase in the number of pronouns without any antecedent. The statistics for each method over the two corpora are presented in Table 3. In addition to these statistics we also computed the ratio between the number of candidates for gender marked pronouns and the number of gender marked pronouns in the text, and the ratio between the the number of antecedents and the number of gender marked pronouns as an indication to how difficult it is to resolve these pronouns. Ideally, a method would reduce the first ratio, leaving the second one unchanged. As can be seen in Table 3 the method which most reduces the number of candidates is the simple method, but it also leads to a dramatic decrease in the number of antecedents available to the anaphora resolution system and a large increase in the number of pronouns without any antecedent. The method which preserves most antecedents is the machine learning method, though it also keeps a large number of candidates.

We note that the two ratios proposed earlier are not very indicative on their own, a formula which combines the number of antecedents, number of candidates and the number of pronouns without any antecedent should be considered instead. Ideally, an animacy filter should eliminate as many candidates as possible without eliminating antecedents, and as a result should not increase the number of pronouns without an antecedent. In light of this, the following formula was used to measure the *imposition*, or degree to which the performance of a filtering system is unwelcome:

$$imposition = \frac{1}{\alpha} * \frac{c_A - a_A}{c_B} + \beta * \frac{a_B - a_A}{a_B} + \gamma * \frac{p_A - p_B}{p_B} \quad (5)$$

where c_B and c_A represent the number of candidates before filtering and after, respectively; a_B and a_A are the

	w/o	Rand.	Wei	Dum	Sim	SL	SN	ML	MLL	MLN	MW	MWL	MWN
Results on the AI Corpus													
Pron.	429	429	429	429	429	429	429	429	429	429	429	429	429
GMP	215	215	215	215	215	215	215	215	215	215	215	215	215
CGMP	3770	2104	2272	1973	2660	2586	2643	2948	2854	2926	2946	2852	2924
Ant	556	369	441	273	405	411	402	540	535	536	528	523	536
$\frac{CGMP}{GMP}$	17.53	9.78	10.56	9.17	12.37	12.02	12.29	13.71	13.27	13.60	13.70	13.26	13.6
$\frac{Ant}{GMP}$	2.58	1.71	2.05	1.26	1.88	1.91	1.86	2.51	2.48	2.49	2.45	2.43	2.49
GMPA	41	74	87	91	53	53	54	42	43	43	45	46	43
Results on the SEMCOR Corpus													
Pron	6006	6006	6006	6006	6006	6006	6006	6006	6006	6006	6006	6006	6006
GMP	3497	3497	3497	3497	3497	3497	3497	3497	3497	3497	3497	3497	3497
CGMP	33360	24790	23786	25604	27881	27749	27876	27246	27100	27223	27231	27084	27207
Ant	5854	5221	5198	5266	5647	5649	5646	5729	5731	5729	5721	5723	5720
$\frac{CGMP}{GMP}$	9.53	7.08	6.80	7.32	7.97	7.93	7.97	7.79	7.74	7.78	7.78	7.74	7.78
$\frac{Ant}{GMP}$	1.67	1.49	1.48	1.50	1.61	1.61	1.61	1.63	1.63	1.63	1.63	1.64	1.64
GMPA	811	1070	1182	1031	890	893	890	861	863	861	863	866	863

Table 3: The results of the extrinsic evaluation (w/o = Performance without filtering on the basis of animacy, Rand. = Performance with random classification of animacy, Wei = Weighted random classification, Dum = Dummy method (NP is inanimate), Sim = Simple method (Evans and Orăsan, 2000), SL = Sim with word lists, SN = Sim with NER, ML = Machine learning method (Orăsan and Evans, 2001), MLL = ML with word lists, MLN = ML with NER, MW = ML with WSD, MWL = ML with WSD and word lists, MWN = ML with WSD and NER, Pron = Number of pronouns in the corpus, Cand = Total number of candidates, GMP = Gender marked pronouns (e.g. he, she, it, etc.), CGMP = Candidates for gender marked pronouns, Ant = total number of antecedents for gender marked pronouns, GMPA = Gender marked pronouns without any antecedent)

number of antecedents before and after filtering; whereas p_B and p_A are the number of pronouns without antecedents before and after filtering. A good method minimises this measure by making each term to tend to zero. The term $c_A - a_A$ indicates the number of candidates which are not antecedents present after the filtering. For a good filtering method the value of this term is small.[6] In order to be able to compare different methods, this value is normalised using the number of candidates present before filtering. Using formula (5), a method which eliminates many antecedents will be penalised by the term $a_B - a_A$ which is again normalised to facilitate comparison. A further penalty occurs when the number of pronouns without an antecedent increases. The three coefficients α, β, γ are weights for each term of the formula which indicate the degree of importance of each one. In our evaluation we used the following values: $\alpha = 1$, $\beta = 2$, $\gamma = 4$ because we considered that the increase in the number of pronouns without an antecedent should be penalised more the elimination of antecedents. Table 4 shows the value of the imposition measure for different methods.

Table 4 reiterates our observation that the machine learning method (presented in Section 2.2.) is the best method for both corpora. The success of this method is closely followed by the systems using machine learning with a gazetteer, and the one using machine learning and a named entity recogniser. As discussed in the next section, the running time is factor which has to be taken into consideration when a system is evaluated. In light of this, the system using a gazetteer should be preferred because it is faster. Surprisingly, the system combining machine learning with word sense disambiguation and named entity recognition performs better than the one with machine learning, word sense disambiguation and a gazetteer, but given the time necessary to run this system and its greater imposition compared with the machine learning method, none of these system is considered a good filter. As expected, all the versions of the system which use the simple method fail to filter the candidates accurately. This can be explained by the low accuracy in identifying animate entities by the simple method. As expected, the

Method	AI	SEMCOR
ML	0.90	1.07
ML+list	0.99	1.07
ML+WSD+NER	1.01	1.08
ML+NER	1.01	1.07
ML+WSD	1.24	1.08
ML+WSD+list	1.33	1.09
Simple+list	2.36	1.27
Simple	2.41	1.26
Simple+NER	2.51	1.26
Random	4.43	2.20
Weighted random	5.47	2.73
Dummy	6.42	2.02

Table 4: The values for *imposition* over the two corpora used

[6] This term cannot be zero because some candidates cannot be eliminated using only gender/animacy agreement constraints

Method	AI	SEMCOR
Simple method	3 sec.	25 sec.
ML	51 sec.	286 sec.
ML+WSD	Several hours	

Table 5: The time necessary for different methods

baseline methods perform significantly worse than all of our methods.

As can be noticed in Table 4, the values of imposition are closer for the SEMCOR corpus, than for the AI corpus. This can be explained by the large number of inanimate entitites in SEMCOR, which makes the filtering easier.

3.3. The complexity of the systems

Another aspect which needs to be considered whenever a system is developed is its complexity. This becomes a very important issue whenever such a system is integrated with a larger system, which needs to react promptly to its input (e.g. systems which are available over the Web). In our case, each method presented in Section 2. is more complex than the previous one, and therefore requires more time to run. Table 5 shows the time necessary to run each system on the two corpora. As can be seen, the fastest method is the *simple method* which has a complexity proportional with *n*m* where *n* is the number of entities in the entire corpus, and *m* is the average number of senses for each word in WordNet. The method which uses machine learning is slower because it has to prepare the data for the machine learning algorithm, a process which has a similar complexity to the simple method, and in addition it has to run the memory-based learning algorithm, which compares each new instance with all instances already seen. Even though TiMBL, the machine learning algorithm used, employs some complex indexing techniques to speed up the process, for large training sets, the algorithm is slow. When word sense disambiguation is used, the processing time increases dramatically, because the complexity of the algorithm used is n^m where n is the number of distinct nouns from a text to be disambiguated, and m is the average number of senses from WordNet for each noun. As we argued in Section 3.1., a system which uses lists of words and named entity recognition attempts to classify more entities than the methods which do not use them, and therefore we did not compute the accuracy of the systems which incorporate them. For the same reason, we did not record the time necessary to run those systems. In addition, the module which performs named entity recognition is written in Perl, whereas the other modules are written in C++, making it impossible to have a direct comparison between them. However, in terms of complexity, the systems which use word lists have a complexity comparable to the corresponding systems which do not. The complexity of the system increases dramatically, and as a result of this so does the running time, when the named entity recogniser is used.

In addition to the information which a module added to a system brings, one has to consider the errors that it introduces. In some cases, it is possible that the additional information brought to the system is not enough to compensate for the errors introduced by the same module, and as a result the overall accuracy of the system decreases instead of increasing. Such a phenomenon was noted in our evaluation. As shown in Sections 3.1. and 3.2. the best system is the one which used machine learning. The other systems which use different modules to enhance the success of the machine learning method, have performance slightly worse than those without them, which indicates that the errors introduced by these modules is greater than the information added to the system. A criticism could be that the word sense disambiguation method used (Resnik, 1995) is rather old and newer methods may perform better. We chose this one because it was very easy to implement and it does not need annotated resources for training. In the future we plan to try other word sense disambiguation methods. The named entity recogniser is still under development and we are confident that later versions will lead to further improvement.

4. Related work

There are two main threads of related work in the literature - the first pertaining to automatic recognition of persons and animate entities, the second pertaining to modular system design and its implications for performance. With regard to work concerned with recognition of NP animacy, we are only concerned with those methods which tackle the problem in English texts, a problem concerned with semantics which cannot be addressed using morphological information, as in other languages.

Identification of the specific gender of proper names has been attempted in work presented in (Hale and Charniak, 1998). That method works by processing a 93931-word portion of the Penn-Treebank corpus with a pronoun resolution system and then noting the frequencies with which particular proper nouns are identified as the antecedents of feminine or masculine pronouns. One of the only articles reporting evaluation results in this area, an accuracy of 68.15% in assigning the correct gender to proper names is reported. Many researchers will be interested in processing new texts of variable lengths and the method presented here will be unsuitable for obtaining relevant information from smaller documents.

WordNet has been used to identify NP animacy in work by (Denber, 1998) and (Cardie and Wagstaff, 1999). Unfortunately, no evaluation of the task of animacy recognition was reported in those papers.

The other issue addressed in this paper concerns the combination of methods to improve the performace of a system. The first attempt was by Edmunson (1969) who combined different sentence extraction methods in an automatic summarisation system. Using an annotated corpus he evaluated different combinations of the modules, and determined which of them leads to the best results. More recently, modules were combined to improve the results of named entity recognition (Mikheev et al., 1999), question answering (Harabagiu et al., 2000) and noun phrase extraction (Tjong Kim Sang et al., 2000). Banko and Brill (2000) proposed an alternative way to improve the

performance of a system. Instead of adding new modules to it, they propose to increase the size of the corpora used for training several times over. They evaluate their approach for confusion set disambiguation.

5. Conclusions

In this paper we assessed different methods for classifying NPs on the basis of their animacy. The systems used were modular and of different levels of complexity. Assessing them under intrinsic and extrinsic methodologies, we found that the complexity of a system was not strongly correlated with its level of performance.

Since we first sought to classify the animacy of NPs in a text in (Evans and Orăsan, 2000) we have realised that a growing number of resources are required in order to meet this challenge. For example, WSD is required in order to accurately obtain the senses of the words in a text given that this information provides the basis for our machine learning method. Initially-capitalised words must be classified using techniques from NER. However, due to the difficulty in obtaining good performance from these modules, their incorporation into the system did not provide the gains in performance that we expected. In fact, their inclusion was detrimental to performance.

From this point, we intend to invest more time and effort in the development of effective methods for WSD and NER.

6. References

Michele Banko and Eric Brill. 2000. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting and the 10th Conference of the European Chapter (ACL2000)*, pages 26 – 33, Toulouse, France, July 9 – 11.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 conference on Empirical Methods in NLP and Very Large Corpora (ACL'99)*, pages 82 – 89, Maryland, USA, 20 June - 26 June.

Nancy A. Chinchor, editor. 1998. *Message Understanding Conference Proceedings*. Science Applications International Corporation. US. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. 2000. An empirical investigation of the relation between discourse structure and co-reference. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*, pages 208 – 214, Saarbrucken, Germany, 31 July - 4 August.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2001. Timbl: Tilburg memory based learner, version 4.0, reference guide. Technical Report ILK Technical Report 01-04, Tilburg University.

Michel Denber. 1998. Automatic resolution of anaphora in English. Technical report, Eastman Kodak Co.

H. P. Edmunson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264 – 285, April.

Richard Evans and Constantin Orăsan. 2000. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 154 – 162, Lancaster, UK, 16 – 18 November.

Christiane Fellbaum, editor. 1998. *WordNet: An Eletronic Lexical Database*. The MIT Press.

John Hale and Eugene Charniak. 1998. Getting useful gender statistics from english texts. Technical report, Brown University.

Sanda M. Harabagiu, Mariu A. Paşca, and Steven J. Maiorano. 2000. Experiments with open-domain textual question answering. In *The 18th International Conference on Computational Linguistics (COLING2000)*, pages 292 – 298, Saarbrucken, Germany, 31 July - 4 August.

Shari Landes, Claudia Leacock, and Randee I. Tengi. 1998. Building semantic concordances. In Fellbaum (Fellbaum, 1998), pages 199 – 216.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of Association for Computational Linguistics*, pages 1 – 9, Bergen, Norway, 8 - 12 June.

Andrei Mikheev. 2000. Document centered approach to text normalization. In *Proceedings of the SIGIR*, pages 136 – 143. ACM, June.

Ruslan Mitkov, Richard Evans, and Constantin Orasan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 168 – 186, Mexico City, Mexico, February 17 - 23. Springer.

Ruslan Mitkov. 2002. *Anaphora resolution*. Longman.

Constantin Orăsan and Richard Evans. 2001. Learning to identify animate references. In *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*, pages 129 – 136, Toulouse, France, July 6 - 7.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI*, pages 448–453.

Karen Sparck-Jones and Julia R. Galliers. 1996. *Evaluating natural language processing systems: an analysis and review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

Erik F. Tjong Kim Sang, Walter Daelemens, Hervé Déjean, Rob Koeling, Yuval Krymolowkki, Vasin Punyakanok, and Dan Roth. 2000. Applying system combination to base noun phrase identification. In *The 18th International Conference on Computational Linguistics (COLING2000)*, pages 857 – 863, Saarbrucken, Germany, 31 July - 4 August.