GETALP: Propagation of a Lesk Measure through an Ant Colony Algorithm

Didier Schwab, Andon Tchechmedjiev, Jérôme Goulian,
Mohammad Nasiruddin, Gilles Sérasset, Hervé Blanchon
LIG-GETALP
Univ. Grenoble Alpes
http://getalp.imag.fr/WSD
firstname.lastname@imag.fr

Abstract

This article presents the GETALP system for the participation to SemEval-2013 Task 12, based on an adaptation of the Lesk measure propagated through an Ant Colony Algorithm, that yielded good results on the corpus of Semeval 2007 Task 7 (WordNet 2.1) as well as the trial data for Task 12 SemEval 2013 (BabelNet 1.0). We approach the parameter estimation to our algorithm from two perspectives: edogenous estimation where we maximised the sum the local Lesk scores; exogenous estimation where we maximised the F1 score on trial data. We proposed three runs of out system, exogenous estimation with BabelNet 1.1.1 synset id annotations, endogenous estimation with BabelNet 1.1.1 synset id annotations and endogenous estimation with WordNet 3.1 sense keys. A bug in our implementation led to incorrect results and here, we present an amended version thereof. Our system arrived third on this task and a more fine grained analysis of our results reveals that the algorithms performs best on general domain texts with as little named entities as possible. The presence of many named entities leads the performance of the system to plummet greatly.

1 Introduction

Out team is mainly interested in Word Sense Disambiguation (WSD) based on semantic similarity measures. This approach to WSD is based on a local algorithm and a global algorithm. The local algorithm corresponds to a semantic similarity measure (for example (Wu and Palmer, 1994), (Resnik, 1995) or (Lesk, 1986)), while the global algorithm propagates the values resulting from these measures at the level of a text, in order to disambiguate the words that compose it. For two years, now, our team has focussed on researching global algorithms. The local algorithm we use, a variant of the Lesk algorithm that we have evaluated with several global algorithms (Simulated Annealing (SA), Genetic Algorithms (GA) and Ant Colony Algorithms (ACA)) (Schwab et al., 2012; Schwab et al., 2013), has shown its robustness with WordNet 3.0. For the present campaign, we chose to work with an ant colony based global algorithms that has proven its efficiency (Schwab et al., 2012; Tchechmedjiev et al., 2012).

Presently, for this SemEval 2013 Task 12 (Navigli et al., 2013), the objective is to disambiguate a set of target words (nouns) in a corpus of 13 texts in 5 Languages (English, French, German, Italian, Spanish) by providing, for each sense the appropriate sense labels. The evaluation of the answers is performed by comparing them to a gold standard annotation of the corpus in all 5 languages using three possible sense inventories and thus sense tags: BabelNet 1.1.1 Synset ids (Navigli and Pozetto, 2012), Wikipedia page names and Wordnet sense keys (Miller, 1995).

Our ant colony algorithm is a stochastic algorithm that has several parameters that need to be selected and tuned. Choosing the values of the parameters based on linguistic criteria remains an open and difficult problem, which is why we wanted to automatize the parameter search process. There are two ways to go about this process: exogenous estima-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 232–240, Atlanta, Georgia, June 14-15, 2013.

tion, when the parameter values are selected so as to maximise the F-score on a small training annotated corpus and then used to disambiguate another corpus (weakly supervised); endogenous estimation, when the parameters are chosen so as to maximise the global similarity score on a text or corpus (unsupervised). Our first experiment and system run consists in tuning the parameters on the trial corpus of the campaign and running the system with the BabelNet sense inventory. Our second and third experiments consist in endogenous parameter estimation, the first using BabelNet as a sense inventory and the second using WordNet. Unfortunately, the presence of an implementation issue prevented us from obtaining scores up to par with the potential of our system and thus we will present indicative results of the performance of the system after the implementation issue was fixed.

2 The GETALP System: Propagation of a Lesk Measure through an Ant Colony Algorithm

In this section we will first describe the local algorithm we used, followed by a quick overview of global algorithms and our own Ant Colony Algorithm.

2.1 The Local Algorithm: a Lesk Measure

Our local algorithm is a variant of the Lesk Algorithm (Lesk, 1986). Proposed more than 25 years ago, it is simple, only requires a dictionary and no training. The score given to a sense pair is the number of common words (space separated strings) in the definition of the senses, without taking into account neither the word order in the definitions (bag-of-words approach), nor any syntactic or morphological information. Variants of this algorithm are still today among the best on English-language texts (Ponzetto and Navigli, 2010).

Our local algorithm exploits the links provided by WordNet: it considers not only the definition of a sense but also the definitions of the linked senses (using all the semantic relations for WordNet, most of them for BabelNet) following (Banerjee and Pedersen, 2002), henceforth referred as $ExtLesk$[1] Contrarily to Banerjee, however, we do not consider the sum of squared sub-string overlaps, but merely a bag-of-words overlap that allows us to generate a dictionary from WordNet, where each word contained in any of the word sense definitions is indexed by a unique integer and where each resulting definition is sorted. Thus we are able to lower the computational complexity from $O(mn)$ to $O(m)$, where m and n are the respective length of two definitions and $m \geq n$. For example for the definition: "Some kind of evergreen tree", if we say that *Some* is indexed by 123, *kind* by 14, *evergreen* by 34, and tree by 90, then the indexed representation is $\{14, 34, 90, 123\}$.

2.2 Global Algorithm : Ant Colony Algorithm

We will first review the principles pertaining to global algorithms and then a more detailed account of our Ant Colony algorithm.

2.2.1 Global algorithms, Global scores and Configurations

A global algorithm is a method that allows to propagate a local measure to a whole text in order to assign a sense label to each word. In the similarity-based WSD perspective, the algorithms require some *fitness* measure to evaluate how good a configuration is. With this in mind, the score of the selected sense of a word can be expressed as the sum of the local scores between that sense and the selected senses of all the other words of a context. Hence, in order to obtain a *fitness* value (*global score*) for the whole configuration, it is possible to simply sum the scores for all selected senses of the words of the context: $Score(C) = \sum_{i=1}^{m} \sum_{j=i}^{m} ExtLesk(w_{i,C[i]}, w_{j,C[j]})$.

For a given text, the chosen configuration is the one which maximizes the global score among the evaluated ones. The simplest approach is the exhaustive evaluation of sense combinations (BF), used for example in (Banerjee and Pedersen, 2002), that assigns a score to each word sense combination in a given context (window or whole text) and selects the one with the highest score. The main issue with this approach is that it leads to a combi-

[1] All dictionaries and Java implementations of all algorithms of our team can be found on our WSD page `http://getalp.imag.fr/WSD` and more specifically for SemEval 2013 Task 12 on the following page `http://getalp.imag.fr/static/wsd/GETALP-WSD-ACA/`

natorial explosion in the length of the context window or text. The number of combinations is indeed $\prod_{i=1}^{|T|}(|s(w_i)|)$, where $s(w_i)$ is the set of possible senses of word i of a text T. For this reason it is very difficult to use the BF approach on an analysis window larger than a few words. In our work, we consider the whole text as context. In this perspective, we studied several methods to overcome the combinatorial explosion problem.

2.2.2 Complete and Incomplete Approaches

Several approximation methods can be used in order to overcome the combinatorial explosion issue. On the one hand, *complete approaches* try to reduce dimensionality using pruning techniques and sense selection heuristics. Some examples include: (Hirst and St-Onge, 1998), based on lexical chains that restrict the possible sense combinations by imposing constraints on the succession of relations in a taxonomy (e.g. WordNet); or (Gelbukh et al., 2005) that review general pruning techniques for Lesk-based algorithms; or yet (Brody and Lapata, 2008) who exploit distributional similarity measures extracted from corpora (information content).

On the other hand, *incomplete approaches* generally use stochastic sampling techniques to reach a local maximum by exploring as little as necessary of the search space. Our present work focuses on such approaches. Furthermore, we can distinguish two possible variants:

- local neighbourhood-based approaches (new configurations are created from existing configurations) among which are some approaches from artificial intelligence such as genetic algorithms or optimization methods such as simulated annealing;
- constructive approaches (new configurations are generated by iteratively adding new elements of solutions to the configuration under construction), among which are for example ant colony algorithms.

2.2.3 Principle of our Ant Colony Algorithm

In this section, we briefly describe out Ant Colony Algorithm so as to give a general idea of how it operates. However, readers are strongly encouraged to read the detailed papers (Schwab et al., 2012; Schwab et al., 2013) for a more detailed description of the system, including examples of how the graph is built, of how the algorithm operates step by step as well all pseudo code listing.

Ant colony algorithms (ACA) are inspired from nature through observations of ant social behavior. Indeed, these insects have the ability to collectively find the shortest path between their nest and a source of food (energy). It has been demonstrated that cooperation inside an ant colony is self-organised and allows the colony to solve complex problems. The environment is usually represented by a graph, in which virtual ants exploit pheromone trails deposited by others, or pseudo-randomly explore the graph. ACAs are a good alternative for the resolution of optimization problems that can be encoded as graphs and allow for a fast and efficient exploration on par with other search heuristics. The main advantage of ACAs lies in their high adaptivity to dynamically changing environments. Readers can refer to (Dorigo and Stützle, 2004) or (Monmarché, 2010) for a state of the art.

In this article we use a simple hierarchical graph (text, sentence, word) that matches the structure of the text and that exploits no external linguistic information. In this graph we distinguish two types of nodes: nests and plain nodes. Following (Schwab et al., 2012), each possible word sense is associated to a nest. Nests produce ants that move in the graph in order to find energy and bring it back to their mother nest: the more energy is brought back by ants, the more ants can be produced by the nest in turn. Ants carry an odour (a vector) that contains the words of the definition of the sense of its mother nest. From the point of view of an ant, a node can be: (1) *its mother nest*, where it was born; (2) *an enemy nest* that corresponds to another sense of the same word; (3) *a potential friend nest*: any other nest; (4) a *plain node*: any node that is not a nest. Furthermore, to each plain node is also associated an odour vector of a fixed length that is initially empty.

Ant movement is function of the scores given by the local algorithm, of the presence of energy, of the passage of other ants (when passing on an edge ants leave a pheromone trail that evaporates over time) and of the nodes' odour vectors (ants deposit a part of their odour on the nodes they go through). When an ant arrives onto the nest of another word (that corresponds to a sense thereof), it can either continue its

exploration or, depending on the score between this nest and its mother nest, decide to build a bridge between them and to follow it home. Bridges behave like normal edges except that if at any given time the concentration of pheromone reaches 0, the bridge collapses. Depending on the lexical information present and the structure of the graph, ants will favor following bridges between more closely related senses. Thus, the more closely related the senses of the nests are, the more bridges between them will contribute to their mutual reinforcement and to the sharing of resources between them (thus forming *meta-nests*); while the bridges between more distant senses will tend to fade away. We are thus able to build interpretative paths (possible interpretations of the text) through emergent behaviour and to suppress the need to use a complete graph that includes all the links between the senses from the start (as is usually the case with classical graph-based optimisation approaches).

Through the emergence of interpretative paths, sense pairs that are closer semantically benefit from an increased ant traffic and thus tend to capture most of the energy of the system at a faster pace, thus favouring a faster convergence over an algorithm that uses a local neighbourhood graph (nodes are senses interconnected so as to represent all sense combinations in a context window) without sacrificing the quality of the results.

The selected answers correspond, for each word to the nest node with the highest energy value. The reason for this choice over using the pheromone concentration is that empirically, the energy level better correlates with the actual F1 scores. In turn, the global Lesk score of a selected sense combination correlates even better with the F1 score, which is why, we keep the sense combinations resulting from each iteration of the algorithm (highest energy nests at each iteration) and select the one with the highest global Lesk score as the final solution.

2.3 Parameters

This version of our ant algorithm has seven parameters (ω, E_a, E_{max}, E_0, δ_v, δ, L_V) which have an influence on the emergent phenomena in the system:

- The maximum amount of energy an ant can carry, E_{max} and E_a the amount of energy an ant can take on a node, influences how much an ant explores the environment. Ants cannot go back through an edge they just crossed and have to make circuits to come back to their nest (if the ant does not die before that). The size of the circuits depend on the moment the ants switch to return mode, hence on E_{max}.
- The evaporation rate of the pheromone between cycles (δ) is one of the memories of the system. The higher the rate is, the least the trails from previous ants are given importance and the faster interpretative paths have to be confirmed (passed on) by new ants in order not to be forgotten by the system.
- The initial amount of energy per node (E_0) and the ant life-span (ω) influence the number of ants that can be produced and therefore the probability of reinforcing less likely paths.
- The odour vector length (L_v) and the proportion of odour components deposited by an ant on a plain node (δ_V) are two dependent parameters that influence the global system memory. The higher the length of the vector, the longer the memory of the passage of an ant is kept. On the other hand, the proportion of odour components deposited has the opposite effect.

Given the lack of an analytical way of determining the optimal parameters of the ant colony algorithm, they have to be estimated experimentally, which is detailed in the following section.

3 Acquisition of Parameter Values

The algorithms we are interested in have a certain number of parameters that need tuning in order to obtain the best possible score on the evaluation corpus. There are three possible approaches:

- Make an educated guess about the value ranges based on *a priori* knowledge about the dynamics of the algorithm;
- Test manually (or semi-manually) several combinations of parameters that *appear* promising and determine the influence of making small adjustments to the values ;
- Use a learning algorithm to automate acquisition of parameters values. We present that approach in the following part.

3.1 Automated Parameter Estimation

Two methods can be used to automatically acquire parameters. The first one consists in maximizing the F-score on an sense-annotated corpus (weak approach) while the second one consist in maximizing the global Lesk score (unsupervised approach).

3.1.1 Generalities

Both approaches are based on the same principle (Tchechmedjiev et al., 2012). We use a simulated annealing algorithm (Laarhoven and Aarts, 1987) combined with a non-parametric statistical (Mann-Whitney-U test (Mann and Whitney, 1947)) test with a p-value adapted for multiple comparisons through False Discovery Rate control (FDR) (Benjamini and Hochberg, 1995). The estimation algorithm operates on all the parameters of the ant colony algorithm described above and attempts to maximise the objective function (Global score, F1). The reason why we need to use a statistical test and FDR rather than using the standard SA algorithm, is that the Ant Colony Algorithm is stochastic in nature and requires tuning to be performed over the distribution of possible answers for a given set of parameter values. Indeed, there is no guarantee that the value resulting from one execution is representative at all of the distribution. The exact nature of the distribution of answers is unknown and thus we take a sampling of the distribution as precise as can be afforded. Thus, we require the statistical test to ascertain the significance between the scores for two parameter configurations.

3.1.2 Exogenous parameter tuning

If we have a sense-annotated corpus at our disposal, it is possible to directly use the F1 value obtained by the system on this reference to tune the parameters of the systems so as to maximise said F1 score. The main issues that arise from such methods are the fact that gold standards are expensive to produce and that there is no guarantee on the generality of the contents of the gold standard. Thus, in languages with little resources we may be unable to obtain a gold standard and in the case one is available, there is a potentially strong risk of over fitting. Furthermore due to the nature of the training, taking training samples in a random order for cross-validation becomes tricky. This is why we also want to test another method that can tune the parameters without using labelled examples. For the evaluation, we estimated parameters on the F1 score on the test corpus for English and French (the only ones available). We used the parameters estimated for English for our English results for our first system run `GETALP-BN1` and the French parameters for the results on French, German, Italian, Spanish.

For English we found: $\omega = 26$, $E_a = 14$, $E_{max} = 3$, $E_0 = 34$, $\delta_v = 0.9775$, $\delta = 0.3577$, $L_V = 25$.

For French: $\omega = 19$, $E_a = 9$, $E_{max} = 3$, $E_0 = 32$, $\delta_v = 0.9775$, $\delta = 0.3577$, $L_V = 25$.

3.1.3 Endogenous parameter tuning

In the context of the evaluation campaign, the absence of an example gold standard on the same version of the resource (synset id mismatch between BabelNet 1.0 and 1.1.1 [2]) made dubious the prospect of using parameters estimated from a gold standard. Consequently, we set out to investigate the relation between the F1 score of the gold standard and the Global Lesk Score of successive solutions throughout the execution of the algorithm.

We observed that the Lesk score is highly correlated to the F1 score and can be used as an estimator thereof. The main quality criterion being the discriminativeness of the Lesk score compared to the F1 score (average ratio between the number of possible F1 score values for a single Lesk score value), for which the correlation is a possible indicator. We make the hypothesis based on the correlation that for a given specific local measure, the global score will be an adequate estimator of the F1 score. Our second system run `GETALP-WSD-BN2` is based on the endogenous parameter estimation. We will not list all the parameters here, as there is a different set of parameters for each text and each language.

3.2 Voting

In previous experiment, as can be expected, we have observed a consistent rise the F1 score when applying a majority vote method on the output of several executions (Schwab et al., 2012). Consequently we followed the same process here, and for all the runs of our system we performed 100 executions and applied a majority vote (For each word, our of all se-

[2] `http://lcl.uniroma1.it/babelnet/`

lected senses, take the one that has been selected the most over all the executions) on all 100 answer files. The result of this process is a single answer file and comes with the advantage of greatly reducing the variability of the answers. Say this voting process is repeated over and over again 100 times, then the standard deviation of F1 scores around the mean is much smaller. Thus, we also have a good solutions to the problem of selecting the answer that yields the highest score, without actually having access to the gold standard.

4 Runs for SemEval 2013 task 12

In this section we will describe the various runs we performed in the context of Task 12. We will first present our methodologies relating to the BabelNet tagged gold standard followed by the methodologies relating to the WordNet tagged gold standard.

4.1 BabelNet Gold Standard Evaluation

In the context of the BabelNet gold standard evaluation, we need to tag the words of the corpus with BabelNet synset ids. Due to the slow speed of retrieving Babel synsets and extracting glosses, especially in the context of our extended Lesk Approach, we pre-generate a dictionary for each language that contains entries for each word of the corpus and then for each possible sense (as per BabelNet). In the short time allotted for the competition, we restrict ourselves to building dictionaries only for the words of the corpus, but the process described can be applied to pre-generate a dictionary for the whole of BabelNet.

Each BabelNet synset for a word is considered as a possible sense in the dictionary. For each synset we retrieve the Babel senses and retain the ones that are in the appropriate language. Then, we retrieve the Glosses corresponding to each selected sense and combine them in as the definition corresponding to that particular BabelNet synset. Furthermore, we also retrieve certain of the related synsets and repeat the same process so as to add the related definitions to the BabelNet synset being considered. In our experiments on the test corpus, we determined that what worked best (i.e. English and French) was to use only relations coming from WordNet, all the while excluding the `r`, `gdis`, `gmono` relation added by BabelNet. We observed a similar increase in disambiguation quality with the Degree (Navigli and Lapata, 2010) algorithm implementation that comes with BabelNet. The `r` relation correspond to the relations in BabelNet extracted from Wikipedia, whereas `gdis` and `gmono` corresponds to relation created using a disambiguation algorithm (respectively for monosemous and polysemous words).

4.2 WordNet Gold Standard Evaluation

In the context of the WordNet gold standard evaluation, we initially thought the purpose would be to annotate the corpus in all five languages with WordNet sense keys through alignments extracted from BabelNet. As a consequence, we exploited BabelNet as a resource, merely obtaining WordNet sense keys through the *main senses* expressed in BabelNet, that correspond to WordNet synsets. Although we were able to produce annotations for all languages, as it turns out, the WordNet evaluation was merely aimed at evaluating monolingual systems that do not support BabelNet at all. For reference, we subsequently generated a dictionary from WordNet only, to gauge the performance of our system on the evaluation as intended by the organisers.

5 Results

We will first present the general results pertaining to Task 12, followed by a more detailed analysis on a text by text basis, as well as the comparison with results obtained on the Semeval 2007 WSD task in terms of specific parts of speech.

5.1 General Results for Semeval-2013 Task 12

Important: implementation issue during the evaluation period During the evaluation period, we had an implementation issue, where a parameter that limited the size of definition was not disabled properly. As a consequence, when we experimented to determine the appropriate relations to consider for the context expansion of the glosses, we arrived at the experimental conclusion that using all relations worked best. However, since it was already the case with WordNet (Schwab et al., 2011), we readily accepted that our experimental conclusion was indeed correct. The issue was indirectly resolved as an unforeseen side effect of another hot-fix applied shortly before the start of the evaluation period.

Given that we were not aware of the presence of a limitation on the definition length before the hot-fix, we performed all the experiments under an incorrect hypothesis which led us to an incorrect conclusion, that itself led to the results we obtained for the campaign. Indeed, with no restrictions on the size of the definition, our official results for this task were consistently inferior to the random baseline across the board. After a thorough analysis of our runs we observed that the sum of local measures (global lesk score) that correlated inversely with the gold standard F1 score, the opposite of what it should have been. We immediately located and corrected this bug when we realized what had caused these bad results that did not correspond at all with what we obtained on the test corpus. After the fix, we strictly ran the same experiment without exploiting the gold standard, so as to obtain the results we would have obtained had the bug not been present in the first place.

Run	Lang.	P	R	F1	MFS
BN1	EN	58.3	58.3	58.3	65.6
	FR	48.3	48.2	48.3	50.1
	DE	52.3	52.3	52.3	68.6
	ES	57.6	57.6	57.6	64.4
	IT	52.6	52.5	52.6	57.2
BN2	EN	56.8	56.8	56.8	65.6
	FR	48.3	48.2	48.3	50.1
	DE	51.9	51.9	51.9	68.6
	ES	57.8	57.8	57.8	64.4
	IT	52.8	52.8	52.8	57.2
WN1	EN	51.4	51.4	51.4	63.0

Table 1: Results after fixing the implementation issue for all three of our runs, compared to the Most Frequent Sense baseline (MFS).

We can see in Table 1, that after the removal of the implementation issues, the scores become more competitive and meaningful compared to the other system, although we remain third of the evaluation campaign. We can observe that there is no large difference between the exogenous results (using a small annotated corpus) and endogenous results. Except for the English corpus where there is a 2% increase. The endogenous estimation, since it is performed on a text by text basis is much slower and resource consuming. Given that the exogenous estimation offers slightly better results and that it requires very little annotated data, we can conclude that in most cases the exogenous estimation will be much faster to obtain.

5.2 A more detailed analysis

In this section we will first make a more detailed analysis for each text on the English corpus, by looking where our algorithm performed best. We restrict ourselves on one language for this analysis for the sake of brevity. As we can see in Table 2, the results can vary greatly depending on the text (within a twofold range). The system consistently performs better on texts from the general domain (T 4, 6, 10), often beating the first sense baseline. For more specialized texts, however, (T 2, 7, 8, 11, 12, 13) the algorithm performs notably lower than the baseline. The one instance where the algorithm truly fails, is when the text in question contains many ambiguous entities. Indeed for text 7, which is about football, many of the instance words to disambiguate are the names of players and of clubs. Intuitively, this behaviour is understandable and can be mainly attributed to the local Lesk algorithm. Since we use glosses from the resource, that mostly remain in the general domain, a better performance in matching texts is likely. As for named entities, the Lesk algorithm is mainly meant to capture the similarity between concepts and it is much more difficult to differentiate two football players from a definition over concepts (often more general).

To further outline the strength of our approach, we need to look back further at a setting with all parts of speech being considered, namely Task 7 from SemEval 2007. As can be seen in Table 3, even though for adjectives and adverbs the system is slightly below the MFS (respectively), it has a good performance compared to graph based WSD approaches that would be hindered by the lack of taxonomical relations. For verbs the performance is lower as is consistently observed with WSD algorithms due to the high degree of polysemy of verbs. For example, in the case of Degree (Navigli and Pozetto, 2012), nouns are the part of speech for which the system performs the best, while the scores for other parts of speech are somewhat lower. Thus, we can hypoth-

Text	Descr.	Len.	F1	MFS	Diff.
1	Gen. Env.	228	61.4	68.9	-7.5
2	T. Polit.	84	51.2	66.7	-15.5
3	T. Econ.	84	52.4	56.0	- 3.6
4	News. Gen.	119	58.8	58.0	0.8
5	T. Econ.	74	39.2	36.5	2.7
6	Web Gen.	210	67.1	64.3	2.8
7	T. Sport.	190	34.2	60.5	-26.3
8	Sci.	153	63.4	67.3	-3.9
9	Geo. Econ.	190	63.2	74.2	-11
10	Gen. Law.	160	61.9	61.9	0
11	T. Sport.	125	56.8	64.0	-7.2
12	T. Polit.	185	64.3	73.0	-8.7
13	T. Econ.	130	68.5	72.6	-4.1

Table 2: Text by text F1 scores compared to the MFS baseline for the English corpus (T.= Translated, Gen.= General, Env.= Environment, Polit.= Politics, Econ.= Economics, Web= Internet, Sport.= Sports, Geo.= Geopolitics, Sci.= Science).

A	P.O.S.	F1	MFS F1	Diff
1108	Noun	79.42	77.4	+1.99
591	Verb	74.78	75.3	-0.51
362	Adj.	82.66	84.3	-1.59
208	Adv.	86.95	87.5	-0.55
2269	All	79.42	78.9	+0.53

Table 3: Detailed breakdown of F1 score per part of speech category for Semeval-2007 Task 7, over results resulting from a vote over 100 executions

esise that using a different local measure depending on the part of speech may constitute an interesting development while allowing a return to a more general all-words WSD task where all parts of speech are considered, even when the resource does not offer taxonomical relation for the said parts of speech.

6 Conclusions & Perspectives

In this paper, we present a method based on a Lesk inspired local algorithm and a global algorithm based on ant colony optimisation. An endogenous version (parameter estimation based on the maximisation of the F-score on an annotated corpus) and an exogenous version (parameter estimation based on the maximisation of the global Lesk score on the corpus) of the latter algorithm do not exhibit a significant difference in terms of the F-score of the result. After a more detailed analysis on a text by text basis, we found that the algorithm performs best on general domain texts with as little named entities as possible (around or above the MFS baseline). For texts of more specialized domain the algorithm consistently performs below the MFS baseline, and for texts with many named entities, the performance plummets greatly slightly above the level of a random selection. We also show that with our Lesk measure the system is best suited for WSD in a more general setting with all parts of speech, however in the context of just nouns, it is not the most suitable local measure. As we have seen from the other systems, graph based local measures may be the appropriate answer to reach the level of the best systems on this task, however it is important not to dismiss the potential of other approaches. The quality of the results depend on the global algorithm, however they are also strongly bounded by the local measure considered. Our team, is headed towards investigating local semantic similarity measures and towards exploiting multilingual features so as to improve the disambiguation quality.

7 Acknowledgements

The work presented in this paper was conducted in the context of the Formicae project funded by the University Grenoble 2 (Université Pierre Mendès France) and the Videosense project, funded by the French National Research Agency (ANR) under its CONTINT 2009 programme (grant ANR-09-CORD-026).

References

[Banerjee and Pedersen2002] Satanjee Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing 2002*, Mexico City, February.

[Benjamini and Hochberg1995] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Brody and Lapata2008] Samuel Brody and Mirella Lapata. 2008. Good neighbors make good senses: Exploiting distributional similarity for unsupervised

WSD. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 65–72, Manchester, UK.

[Dorigo and Stützle2004] Dorigo and Stützle. 2004. *Ant Colony Optimization*. MIT-Press.

[Gelbukh et al.2005] Alexander Gelbukh, Grigori Sidorov, and Sang-Yong Han. 2005. On some optimization heuristics for Lesk-like WSD algorithms. In *International Conference on Applications of Natural Language to Information Systems – NLDB'05*, pages 402–405, Alicante, Spain.

[Hirst and St-Onge1998] G. Hirst and David D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic Lexical Database. C. Fellbaum. Ed. MIT Press. Cambridge. MA*, pages 305–332. Ed. MIT Press.

[Laarhoven and Aarts1987] P.J.M. Laarhoven and E.H.L. Aarts. 1987. *Simulated annealing: theory and applications*. Mathematics and its applications. D. Reidel.

[Lesk1986] Michael Lesk. 1986. Automatic sense disambiguation using mrd: how to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC '86*, pages 24–26, New York, NY, USA. ACM.

[Mann and Whitney1947] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

[Miller1995] George A. Miller. 1995. Wordnet: A lexical database. *ACM*, Vol. 38(No. 11):p. 1–41.

[Monmarché2010] N. Monmarché. 2010. *Artificial Ants*. Iste Series. John Wiley & Sons.

[Navigli and Lapata2010] Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:678–692, April.

[Navigli and Pozetto2012] Roberto Navigli and Simone Paolo Pozetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250. http://dx.doi.org/10.1016/j.artint.2012.07.004.

[Navigli et al.2013] Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, Atlanta, Georgia, 14-15 June.

[Ponzetto and Navigli2010] Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531.

[Resnik1995] Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Schwab et al.2011] Didier Schwab, Jérôme Goulian, and Nathan Guillaume. 2011. Désambiguïsation lexicale par propagation de mesures semantiques locales par algorithmes a colonies de fourmis. In *TALN*, Montpellier (France), Juillet.

[Schwab et al.2012] Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev, and Hervé Blanchon. 2012. Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation. In *Proceedings of COLING'2012*, Mumbai (India), December. To be published.

[Schwab et al.2013] Didier Schwab, Jerôme Goulian, and Andon Tchechmedjiev. 2013. Theoretical and empirical comparison of artificial intelligence methods for unsupervised word sense disambiguation. *Int. J. of Web Engineering and Technology*. In Press.

[Tchechmedjiev et al.2012] Andon Tchechmedjiev, Jérôme Goulian, Didier Schwab, and Gilles Sérasset. 2012. Parameter estimation under uncertainty with simulated annealing applied to an ant colony based probabilistic wsd algorithm. In *Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology*, pages 109–124, Mumbai, India, December. The COLING 2012 Organizing Committee.

[Wu and Palmer1994] Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting of Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.