

Evaluation and Comparison of multilingual fusion strategies for similarity-based Word Sense Disambiguation

Andon Tchechmedjiev, Jérôme Goulian, and Didier Schwab

Univ. Grenoble Alpes
Grenoble, France

{andon.tchechmedjiev, jerome.goulian, didier.schwab}@imag.fr
<http://getalp.imag.fr/wsd>

Abstract. In this article, we investigate the effects on the quality of the disambiguation of exploiting multilingual features with a similarity-based WSD system based on an Ant Colony Algorithm. We considered features from one, two, three or four languages in order to quantify the improvement brought by using features from additional languages. Using BabelNet as a multilingual resource, we considered three data fusion strategies: an early fusion strategy, and two late fusion strategies (majority vote and weighted majority vote). We found that the early fusion approach did not produce any significant improvements while voting strategies adding features from more languages led to an increase in the quality of the disambiguation of up to 2.84%. Furthermore, a simple majority vote led to better results than the weighted variant.

Key words: Similarity-based WSD, Multilingual WSD, Multilingual Features, BabelNet, Fusion Strategies

1 Introduction

Word Sense Disambiguation (WSD) is a core problem in Natural Language Processing (NLP), as it may improve many of its applications, such as multilingual information extraction, automatic summarization or machine translation. More specifically, the aim of WSD is to find the appropriate sense(s) of each ambiguous word in a text from a pre-defined sense inventory.

Focused mainly on the disambiguation of English language text, many WSD algorithms exist and can be classified according different typologies. Figure 1 presents various types of WSD methods according to two dimensions: the use of sense-annotated corpora (y axis) and the use of large lexical databases (x axis). The more an approach is located towards the top of the figure, the more it requires sense-annotated corpora; the more an approach is located towards the right of the figure, the more the quality and size of the lexical databases used is important. Furthermore, the more an approach is located in the upper right corner, the more it will lead to good disambiguation results. Building a

good quality large scale lexical resource and large annotated corpora for one specific language is obviously very costly and one can consider that such resources currently only exist for English.

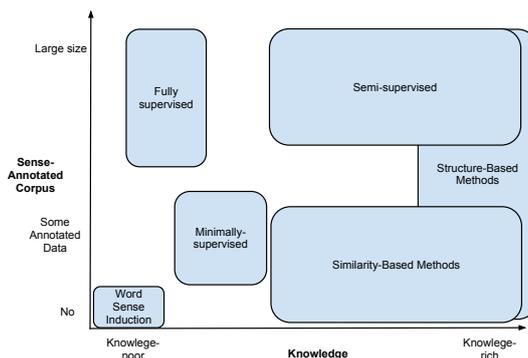


Fig. 1: One WSD approaches classification

In this article, we will focus on similarity based methods. These methods assign scores to word senses through semantic similarity (between word senses), and globally find the sense combinations maximising the score over a text. In other words, a local measure is used to assign a similarity score between two lexical objects (senses, words, constituencies) and the global algorithm is used to propagate the local measures at a higher level.

One of the first knowledge-based method was proposed by [1] and assesses semantic similarity as the number of words in common between sense definitions. The Lesk measure is still widely used today, however, in its original form, suffers from the small size of word-sense definitions. Methods have been proposed to partly solve the issue, either by exploiting the definitions of related senses from a structured lexical resource [2] or by considering the surrounding text in the computation of the overlap [3].

The objective of this article is to evaluate if exploiting multilingual features leads to improved results. We successively experiment adding one, two or three language to our WSD system based on a global ant colony algorithm. In the next sections, we first present a general account of data fusion strategies and then describe the semantic similarity measure and the ant colony algorithm that we use. Subsequently we will describe the method we use to generate both monolingual and combined multilingual dictionaries for our extended Lesk measure (early fusion), followed by a description of the different voting strategies we considered, our experimental protocol and the results.

2 Contribution of multilingual informations for WSD

Using multilingual information for Word Sense Disambiguation can be achieved in two ways: either combining multilingual information before the execution of the disambiguation algorithm (early fusion) or combining multilingual information after the execution of the disambiguation algorithm (late fusion).

The concepts of early and late fusion come from the field of classification and were introduced to deal with the combination of multiple sources of information. The objective of information fusion is thus to improve the performance and robustness of classification systems relying on multi-modal or multiple-source mono-modal input ([4], Chapter 4.3).

2.1 Early and late fusion

One of the first works accounting of a combination of multilingual features using an early fusion approach has been proposed by [5], where a unigram word frequency vectorial representation is used. A monolingual vector for each word sense is first created from an english definition. An MT system (Google translate) is then used to obtain features in French, German and Spanish that are subsequently added to the vectors. Each vector is then weighted with a normal distribution for each of the languages. For the evaluation, combinations of English with one, two or three other languages are considered. The model leads to a reduction in the number of classification errors between 10.58% and 25.96%. Moreover, the more languages are added, the better are the results obtained. More recently, [6] have proposed a similar approach, albeit not with a vectorial model, but derived from a simplified lesk algorithm [3] that selects the word senses that maximise the overlap with the words of the surrounding context rather than between combinations of word-senses. For each word sense, the context in English is translated using an MT system in order to obtain multilingual representations. The purpose is to compare the classification error rate reduction between using only one language (English, and through MT German, French, Spanish) to that of using a concatenated context of all four languages. The evaluation was made on the Semeval 2007 lexical sample task and reports a 26% maximal error rate reduction with four combined languages. Hasan et al. [7] have undertaken to evaluate the influence of adding features from multiple languages, but at the level of semantic relatedness measures using either a MT system or manual translation to generate features. They reach the conclusion that additional information can be acquired from multilingual features, thus significantly improving the correlation with human judgement. However they also find that the improvement is all the more significant for languages with scarce lexical resources (37% improvement for Romanian for example against 8% for English).

Late fusion approaches have mostly been used to combine monolingual features through classifier voting strategies. [8] evaluate the combination of different systems through several late fusion strategies on Senseval-1 English as well as Senseval-2 with 4 languages individually. They report improvements of up to

4% over the best published score for the respective campaigns. [9] have similarly evaluated a broad range of WSD algorithms on the Web-Harvested Corpus Annotated with GermaNet Senses, as well as combinations thereof using several voting strategies (majority vote, weighted majority vote, Borda count), which led to an improvement of 7.23% compared to the best performing standalone algorithm (a variant of Lesk). [10] evaluate two WSD systems, one that uses monolingual features only and one that exploits multilingual information extracted from parallel corpora; their combination through a voting strategy leads to an improvement of 2.24%, which constitutes an indication that multilingual features can indeed be combined beneficially through a vote strategy.

2.2 Motivations: Multilingual Word Sense Disambiguation

The results mentioned in the previous section clearly show that there is a lot of potential in using multilingual informations to improve WSD performance. Our work group currently focuses on multilingual Word Sense Disambiguation. The aim of this task is to find the appropriate sense(s) of each word of a text whatever its language (more reasonably for a certain number of languages). The evaluation requires a reference corpus in various languages. Such an evaluation will be carried out in the SemEval 2013 evaluation campaign (task 12: *Multilingual Word Sense Disambiguation*, march 2013). At the time of the writing, neither the corpus nor the gold standard of the campaign were available. Instead, we used the English all-words task 7 corpus [11] of the SemEval 2007 campaign, where organizers provided a coarse-grained pre-defined sense inventory from WordNet. Considering coarse grained senses distinction means that close senses are counted as equivalent (e.g. snow/precipitation and snow/cover).

3 The LIG-GETALP WSD System

The LIG-GETALP system is a similarity based system. At the local level (between two senses), we use a Lesk similarity measure and at the global level we use our own ant colony algorithm [12–14].

3.1 Local similarity measure: an adaptation of Extended Lesk

Our local algorithm is a variant of the Lesk similarity measure [1], where the score given to a sense pair is the number of words in common between the two definitions, without taking into account either the word order or any syntactic or morphological information.

In this work, we use the definitions from BabelNet, a large scale multilingual lexical resource built from WordNet and Wikipedia (see section 4).

3.2 Global algorithm: Ant Colony Algorithm

Ant colony algorithms (ACA) are inspired from nature through observations of ant social behaviour. 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 allow 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 [15] or [16] 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 [13], 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 (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 favour 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).

4 BabelNet, a large scale multilingual resource

BabelNet [17] is a large scale multilingual lexical resource built from the automatic mapping between WordNet synsets and Wikipedia pages. BabelNet is

based on the concept of a *Babel Synset* that contains a WordNet synset and a set (or sets) of related Wikipedia pages. The latter include Wikipedia pages mapped to WordNet synsets (through a disambiguation algorithm), pages related to the mapped pages (inter-page hyper-links), as well as corresponding multilingual pages obtained through Wikipedia’s inter-language links and their redirections towards pages on the Wikipedia in the respective languages. For concepts that are not covered in all languages supported by BabelNet (English, French, German, Italian, Spanish and Catalan), a set of related sentences are extracted from the text in linked Wikipedia pages and from Semicor instances [18] and are then translated in the missing languages through a machine translation system, thus guaranteeing a homogeneously high coverage. Figure 2 is an illustration of BabelNet’s structure. Additionally, BabelNet provides glosses (definitions) associated to each Babel synset in all the supported languages, which is invaluable useful for Lesk-based approaches.

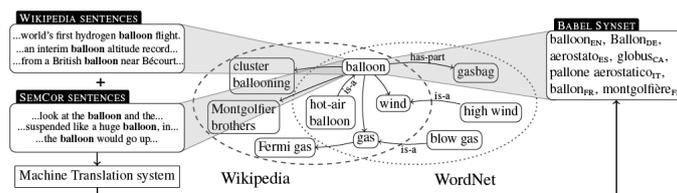


Fig. 2: An overview of BabelNet (figure 2 from [17])

5 Data fusion strategies

As mentioned in previous sections, there are two main data fusion strategies: either combining informations before the execution of the disambiguation algorithm (early fusion) or combining informations after the execution of the disambiguation algorithm (late fusion). In this article, early fusion is achieved by concatenating linked definitions in our lexical resource (BabelNet) and late fusion is achieved by combining the results of the disambiguation obtained for each language. Figure 3 provides illustrations of both strategies for our WSD similarity-based approach.

5.1 Early fusion, *a priori* generation of multilingual dictionaries

We first generate monolingual dictionaries for the four languages chosen for evaluation: English, French, German and Spanish, from which we obtain monolingual results from the execution of our disambiguation algorithm. These results will be used as baselines to evaluate the potential improvement brought by adding more

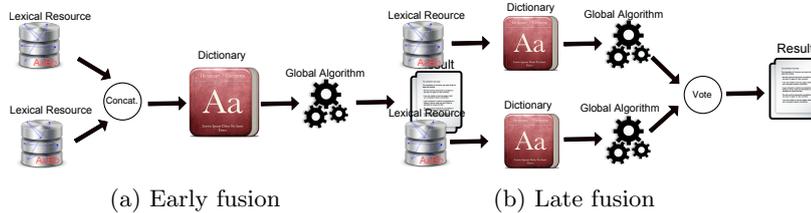


Fig. 3: Illustrations of the fusion strategies for Knowledge-based WSD

languages. For the early fusion, we generated dictionaries for all the combinations of 1, 2, 3 and 4 languages (a total of 14 dictionaries) that were subsequently used for the experiments.

5.2 Late fusion voting strategies

As explained earlier, late fusion corresponds to a combination of the output of different classifiers. In the case of results generated by WSD systems for the Semeval All-words task 7, the output of the algorithm is sense assignments for the whole corpus, which exactly corresponds to the output of a classifier over a set of instances. However, the format of the results of the campaign only requires to provide sense labels for each word, which means we have no information on the confidence in the assignments or any form of posterior probability distribution. As such, only a simple majority voting mechanism is applicable [8]. Of course in the case of our ant colony algorithm, each node corresponding to a word sense has an associated energy level that we use to select the final answer. While it does not constitute a posterior probability distribution as such, it offers likelihood values that can be used as weights for a weighted majority vote.

So as to be able to make a fair comparison with other systems, we considered both a simple majority vote and a weighted majority vote.

Specifically in the case of the fusion of multilingual resources, the vote is performed on result files generated by the execution of the algorithm using each of the corresponding monolingual dictionaries as input. For example, in order to apply a late fusion strategy to combine French and English, one needs to make several executions with the dictionaries in both languages. Subsequently the voting strategy can be applied on the result files (in equal number) of French and English executions, to obtain a single result file.

In our experiments, we used 12 input files in total per vote, distributed equally among the languages. Of course this means that we need 12 times more executions to apply a late vote strategy, which constitutes a significant amount of time. In order to mitigate this problem, over 120 executions, we applied the vote on 12 successive answers by increments of 1 modulo 120 so as to obtain 120 vote results without requiring more executions. Even though the samples are somewhat correlated, we found no evidence of impact on the quality of the results obtained compared to always using separate executions.

For the majority vote strategy, for each word of the corpus, we build an histogram of the cumulative distribution of selected senses among the result files. Subsequently, we select for each word the sense with the relative majority (maximum frequency).

For the weighted vote, instead of giving only a single vote for one sense selection, several votes (the weight values) are given for each sense selection. In the case of our algorithm, we had a different energy value for each sense in each result file that we used as a weight. Like for the unweighted majority vote, for each word, the sense with the maximum score (sum of energy values) is selected.

The next section describes more precisely the experiments and their outcome.

6 Experimental evaluation and results

6.1 Experimental Protocol

The objective of our experiments is to determine if multilingual features do indeed improve the quality of disambiguation, and, if so, whether or not the number of languages have an influence on the degree of improvement. Furthermore, we want to know, which, among early fusion and two late fusion techniques for the combination of features, yields the best and most salient results.

Since our algorithm is stochastic in nature, we need to have a representation of the distribution of solutions as accurate as possible in order to make a statistically significant analysis. Thus, we used 120 samples for each of the groups of one, two, three and four languages. In order to avoid any bias we systematically considered all possible language combinations for 1, 2 or 3 languages and the whole language group (the 4 languages together). Depending on the number of combinations, we made for each combination a number of runs so as to obtain 120 executions in total for the whole language group. For example for the 1 language group, we considered 30 executions of English, French, German and Spanish and grouped them together.

In order to study the distribution of the scores depending on each language group, we used a one-way ANOVA variance analysis, coupled with a Tukey's HSD post-hoc pairwise analysis. The one-way ANOVA analysis is a generalization to several groups of the unpaired t-test that compares the means of only two samples. This method offers a way to test the hypothesis of equality of the mean scores at different levels, e.g between the different groups of languages. If the difference is statistically significant (with a p-value $p < 0.05$), one needs to determine which are the language pairs whose scores are significantly different; Tukey's method is a way to make all pairwise comparisons of mean scores using studentized range distribution to prevent the accumulation of type I errors (incorrect rejection of a true null hypothesis).

6.2 Results

Figure 4 and Tables 1 and 2 present the average scores and standard deviations for: early fusion, the majority vote and weighted majority vote respectively.

Similarly, Figures 4, 5a and 5b, present box plots of the results; the dashed line is the WordNet first sense baseline. The parenthesis with a † lists the language groups with significant differences in the average score (Tukey’s HSD test).

In order to check the normality assumption for ANOVA, we computed the correlation between the theoretical (normal distribution) and the empirical quantiles. For all language combinations there always was a correlation above 0.99. Furthermore we used Levene’s variance homogeneity test and found a minimum significance level of 10^{-6} between all algorithms and metrics.

Before analysing the results, it is important to note that the purpose of this experiment is not to obtain results better than the state of the art, but rather to see the general effect of adding multilingual features without any consideration of a particular language. Indeed, for reference, in a purely monolingual setting, we obtain on average above 77% using a dictionary generated from WordNet only and above 79% using a majority vote strategy. We have noted a decreased score when using only BabelNet compared to using WordNet on English for our WSD algorithm with the current parameters, however for languages other than English, the results remain very interesting. Let us first look at Figure 4 that present the results for the early fusion of multilingual features.

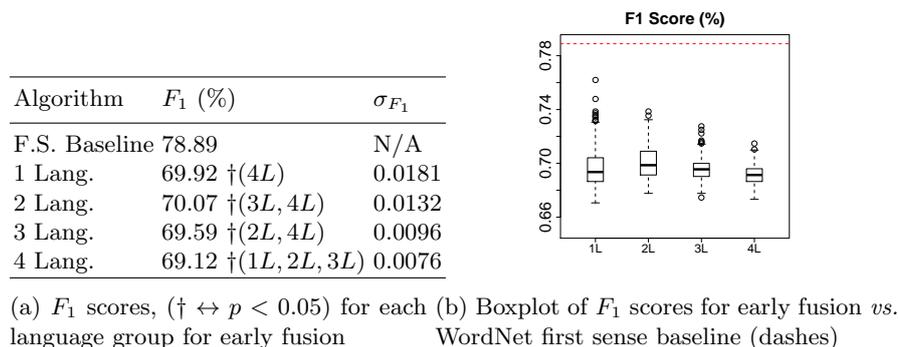


Fig. 4: Results for early fusion

For one language, the average score is quite low, below 70%. However, we can see that the results over English are notably higher than the others (indicated by the many outliers at the top of the distribution). When adding more languages, the general tendency is that the score decreases but also the standard deviation. We can see that the more languages are added, the less outliers there are on the top of the distribution. Furthermore, there are no significant improvements when going from one to two languages. Adding more languages has the effect of causing a slight but significant decrease of the average towards the bottom of the distributions (they roughly share the same minimal value).

We hypothesize that because we only compute our local score as a non-normalised bag of words overlaps and because the number of related *Babel*

synsets can vary, we find ourselves confronted with definitions of very different lengths. Thus, given that longer definitions are privileged, a strong bias is added to the system. Furthermore, given the automatic construction of BabelNet, we think that there may be a higher degree of noise compared to just WordNet, which may very well cause a shift in the parameter space and thus lead to bad scores for parameters that worked much better when only WordNet was used.

For the results of the majority vote strategy, let us look at table 1 and figure 5a. Compared to the one language case with no vote (69.92%), the vote strategy leads to a leap in the score of 4.14%, which is in accordance with the improvements observed when using voting strategies in a monolingual setting [8, 9]. In terms of the improvement when adding more languages, we can see a steady and significant increase in the average score of +1.73% for two languages and +2.32% for three, while a fourth language did not lead to any significant improvements compared to three languages. Furthermore, adding more languages leads to a more compact distribution: between one language and four languages the standard deviation is divided by six.

For the weighted majority vote (table 2 and figure 5b), we can make the same observations about the effect of adding more languages, except that the scores are notably lower than with a simple majority vote. Indeed, for one language, the difference in the average scores between the majority vote and the weighted majority vote is -1.63%, for two languages -1.86%, for three languages -1.45% and for four languages -1.27%. It appears that the majority vote leads to better results in the WordNet monolingual setting. Given that the majority vote is reported as well to show better results by [9], it may be beneficial to consider other simple weighting schemes, for example different weights per part of speech.

Algorithm	F_1 (%)	σ_{F_1}
F.S. Baseline	78.59	N/A
1 Lang.	74.06 †(2L, 3L, 4L)	0.020
2 Lang.	75.79 †(1L, 3L, 4L)	0.0172
3 Lang.	76.38 †(1L, 2L)	0.0112
4 Lang.	76.90 †(1L, 2L)	0.0033

Table 1: Comparison of the F_1 scores, († $\leftrightarrow p < 0.05$) for each language group for majority vote

Algorithm	F_1 (%)	σ_{F_1}
F.S. Baseline	78.59	N/A
1 Lang.	72.43 †(2L, 3L, 4L)	0.0172
2 Lang.	73.94 †(1L, 3L, 4L)	0.0126
3 Lang.	74.93 †(1L, 2L, 4L)	0.0083
4 Lang.	75.63 †(1L, 2L, 3L)	0.0028

Table 2: Comparison of the F_1 scores, († $\leftrightarrow p < 0.01$) for each language group for weighted majority vote

7 Conclusion and perspectives

In this article, we evaluated on the all-words WSD task of Semeval 2007 the effects on disambiguation quality of exploiting multilingual features with our WSD system based on an Ant Colony Algorithm. Multilingual informations are provided by BabelNet, a lexical multilingual resource available since 2012, and built automatically in aligning Wikipedia and WordNet. Three data fusion strategies

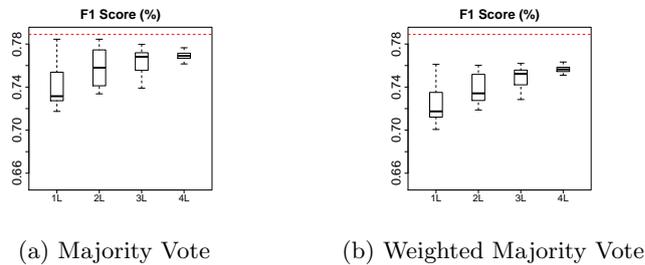


Fig. 5: Box plots of the F_1 scores for majority vote (a) and weighted majority vote (b) compared to the WN FS baseline (dashed line)

have been used: an early fusion strategy, and two vote-base strategies (majority vote and weighted majority vote). We have found that with the early fusion strategy adding multilingual features did not lead to any significant improvements. Integrating information from BabelNet directly lead to poorer scores than with just WordNet. We believe the main cause is that BabelNet is slightly noisier than WordNet. Our ant colony algorithm requiring a minimal level of parameterization, we believe that the parameters that offered very good results with WordNet were far from optimal when using a lexical resource where a significant number of definitions are built with Machine Translation Systems. However, when using late fusion strategies (voting strategies), there was a clear improvement when adding multilingual features from more and more languages. Possible approaches to improve the results we obtained with early fusion would be to consider different ways of concatenating the glosses, for instance imposing a global maximum length for concatenated definitions. Another approach would be to consider the fusion at the feature level, and to use several monolingual dictionaries at a time in our ant colony algorithm by having different types of ant that use features from a different language and to combine the features in the environment through emergence.

Acknowledgments

The work presented in this paper was conducted in the Videosense project, funded by the French National Research Agency (ANR) under its CONTINT 2009 programme (grant ANR-09-CORD-026).

References

1. Lesk, M.: Automatic sense disambiguation using mrd: how to tell a pine cone from an ice cream cone. In: Proceedings of SIGDOC '86, New York, NY, USA, ACM (1986) 24–26
2. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: CICLing 2002, Mexico City (2002)

3. Vasilescu, F., Langlais, P., Lapalme, G.: Evaluating variants of the lesk approach for disambiguating words. In: Proceedings of LREC 2004, Lisbon, Portugal (2004)
4. Bezdek, J., Keller, J., Krishnapuram, R., Pal, N.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers (1999)
5. Banea, C., Mihalcea, R.: Word sense disambiguation with multilingual features. In: Proceedings of ICCS'2011, Oxford, United Kingdom, ACL (2011) 25–34
6. Fernandez-Ordonez, E., Mihalcea, R., Hassan, S.: Unsupervised word sense disambiguation with multilingual representations. In: Proceedings of LREC'2012, Istanbul, Turkey (2012)
7. Hassan, S., Banea, C., Mihalcea, R.: Measuring semantic relatedness using multilingual representations. In: Proceedings of SemEval 2012, Montréal, Canada, ACL (2012) 20–29
8. Florian, R., Yarowsky, D.: Modelling consensus: classifier combination for word sense disambiguation. In: Proceedings of EMNLP '2002, University of Pennsylvania, Philadelphia, ACL (2002) 25–32
9. Henrich, V., Hinrichs, E.: A comparative evaluation of word sense disambiguation algorithms for german. In: Proceedings of LREC'2012, Istanbul, Turkey, ELRA (2012)
10. Guo, W., Diab, M.: Combining orthogonal monolingual and multilingual sources of evidence for all words *wsd*. In: Proceedings of ACL'2010, Uppsala, Sweden (2010) 1542–1551
11. Navigli, R., Litkowski, K.C., Hargraves, O.: Semeval-2007 task 07: Coarse-grained english all-words task. In: SemEval-2007, Prague, Czech Republic (2007) 30–35
12. Schwab, D., Goulian, J., Guillaume, N.: Désambiguïisation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. In: TALN, Montpellier (France) (2011)
13. Schwab, D., Goulian, J., Tchechmedjiev, A., Blanchon, H.: Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation. In: Proceedings of COLING'2012, Mumbai (India) (2012) To be published.
14. Tchechmedjiev, A., Goulian, J., Schwab, D., Sérasset, G.: 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, Mumbai, India, The COLING 2012 Organizing Committee (2012) 109–124
15. Dorigo, Stützle: Ant Colony Optimization. MIT-Press (2004)
16. Monmarché, N.: Artificial Ants. Iste Series. John Wiley & Sons (2010)
17. Navigli, R., Pozetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence (2012) <http://dx.doi.org/10.1016/j.artint.2012.07.004>.
18. Miller, G.A., Leacock, C., Teng, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the workshop on Human Language Technology. HLT '93, Stroudsburg, PA, USA, Association for Computational Linguistics (1993) 303–308