Word Sense Disambiguation in Opinion Mining: Pros and Cons

Tamara Martín-Wanton[1], Alexandra Balahur-Dobrescu[2], Andrés Montoyo-Guijarro[2] and Aurora Pons-Porrata[1]

[1]Universidad de Oriente, Center for Pattern Recognition and Data Mining
Patricio Lumumba s/n, Santiago de Cuba, Cuba
{tamara, aurora}@cerpamid.co.cu
[2]University of Alicante, Departament of Software and Computing Systems,
Apartado de Correos 99, E-03080 Alicante, Spain
{abalahur, montoyo}@dlsi.ua.es

Abstract. The past years have marked the birth of a new type of society - that of interaction and subjective communication, using the mechanisms of the Social Web. As a response to the growth in subjective information, a new task was defined - opinion mining, dealing with its automatic treatment. As the majority of natural language processing tasks, opinion mining is faced with the issue of language ambiguity, as different senses of the same word may have different polarities. This article studies the influence of applying word sense disambiguation (WSD) within the task of opinion mining, evaluating the advantages and disadvantages of the approach. We evaluate the WSD-based method on a corpus of newspaper quotations and compare it to the results of an opinion mining system without WSD. Finally, we discuss our findings and show how WSD helps in the task of opinion mining.

1 Introduction

The past years, with the growing volumen of subjective data originating from texts pertaining to the Social Web -blogs, reviews, forums, discussion panels - have marked the birth of a new type of society, where individuals can freely communicate and exchange opinions. The large benefits that can be obtained by the analysis of this data (more informed customers, companies, societies) made essential the study of methods that can be automatically employed to extract the required information.

Therefore, over the past few years, there has been a large increase of interest in the identification and automatic extraction of the attitudes, opinions and feelings expressed in texts. This movement is given by to the need to provide tools for users of different domains, which require, for different reasons, the automatic monitoring of information that expresses opinion. A system that automatically carries out this task would eliminate the effort to manually extract useful knowledge from the information available on the Web.

Opinion Mining (also known as sentiment classification or subjectivity analysis) covers a wide area of Natural Language Processing, Computational Linguistics and Text Mining. The goal is not to determine at the topic of a document is, but the opinion that is expressed in it. Therefore, its objective is to determine the opinion of a speaker or a writer on a topic [1].

Many approaches to sentiment analysis rely on lexicons of words that may be used to express subjectivity; these works do not make distinction between different senses of a word, so that the term, and not its senses, are classified. Moreover, most subjectivity lexicons are compiled as lists of keywords, rather than word meanings. However, many keywords have both subjective and objective senses and even the purely subjective senses have a degree of positivity or negativity, depending on the context where the corresponding word appears.

This paper presents a research on the advantages of using the word sense disambiguation to determine the polarity of opinions and the role of existing resources. To this end, we evaluated two unsupervised approaches, one based on lists of positive and negative words and the other based on a word sense disambiguation algorithm over the same affect lexicons.

2 Related Work

A major task of Opinion Mining consists in classifying the polarity of the extracted opinion. This process determines whether the opinion is positive, negative or neutral with respect to the entity which it is referring to (for example, a person, a product, a topic, a movie, etc.).

The large majority of research in this field focused on annotating the sentiment of the opinions but out of context (e.g. [2–5]). Recently some works have been published determining the polarity of the word senses, thus building resources that can be useful in different tasks of Opinion Mining ([1, 6, 7]). Esuli and Sebastiani [1] determine the polarity of word senses in WordNet, distinguishing among positive, negative and objective. They manually annotate a seed set of positive/negative senses in WordNet and by following the relations in WordNet expand the small set using a supervised approach. They extend their work [6] by applying the Page Rank algorithm for ranking the WordNet senses in terms of how strongly a sense possesses a given semantic property (e.g., positive or negative). Wiebe and Mihalcea [7] label word senses in WordNet as subjective or objective. They use a method relying on distributional similarity as well as an independent, large manually annotated opinion corpus (MPQA) [8] for determining subjectivity.

Only few recent works take into account the correct senses of the words in the opinions (e.g. [9, 10]), but they are supervised methods. Akkaya et al. [9] build and evaluate a supervised disambiguation system that determines whether a word in a given context is being used with objective or subjective senses. In this approach the inventory of objective and subjective senses of a word can be viewed as an inventory of the senses of the word but with a coarse granularity. Rentoumi et al. [10] go a step further and determine the polarity by disambiguating the words and then mapping the senses to models of positive and negative polarity. To compute these models and produce the mappings of senses, they adopt a graph-based method which takes into account contextual and sub-word information.

3 Motivation and Contribution

Whereas most research concentrates on the analysis of opinions at a word level there are some approaches dealing with the analysis at the sense level. The reason for this fact is that the meaning of most words depends on the context where they appears and in order to determine the polarity of opinions, it is also important to take into account the meanings of the words and the relations between them. The recent studies that regard word senses concentrate on the detection of subjectivity or on ranking senses in a lexicon according to their polarity, but they do not have as main aim the classification of the polarity. On the other hand, most research concentrates on assessing the polarity of opinion using one of the available lexicons of opinion. However, using a resource or another cannot give a measure of the impact the use of these resources has on the final system results.

The main motivation of the present research is to study the impact of word sense disambiguation for determining the polarity of opinions. There are approaches that perform the analysis at the sense level, but they lack of a thorough study of the advantages and disadvantages of the disambiguation as intermediate task. They do not perform the analysis of the approaches at a word and sense levels on the same corpus and using the same resource.

Thus, the contribution of this paper is given by the evaluation of two unsupervised approaches on the same corpus and using the same resources, both are based on knowledge but differ in the level at which the analysis is performed (word and sense level). The second objective is to carry out an analysis of existing public resources for opinion mining and their influence on both approaches. In this way, we can have a clear measure of the impact given by the use of each of the resources separately, as well as study methods to combine them, to obtain better results.

We show that word sense disambiguation avoids the lack of balance between the classification of positive and negative opinions present in an approach as simple as positive or negative words belonging to an opinion. The majority of the existing resources that have annotated senses with its corresponding polarity have little coverage.

4 Experiments and Evaluation

In the experiments, we intend to evaluate the impact of the disambiguation of words in the task of polarity classification of an opinion. With this aim, we first present a "bag of words" approach, and then a second one that uses a word sense disambiguation algorithm to determine the correct sense of the words in the opinion. In both approaches, we firstly perform a pre-processing of the text including sentence recognition, stop-word removal, part-of-speech tagging and word stemming by using the TreeTagger tool [11].

We comparatively analyse the different possible methods and resources for opinion mining that are publicly available and explore the possibility to combine them in order to increase the accuracy of the classification. For the evaluation of both methods we use precision, recall and F1 measures for the Positive (P+, R+, F1+) and Negative (P-, R-, F1-) categories, and the overall precision, recall and F1 (P, R, F1). Additionally, the coverage (Cov) of the method is calculated as the ratio of the number of opinions classified as negative or positive over the total number of opinions.

4.1 Data and Resources

For our experiments, we chose a set of 99 quotes described in [12], on which agreement between a minimum of two annotators could be reached regarding their classification in the positive and negative categories, as well as their being neutral/controversial or improperly extracted. In this paper, we only use the 68 quotes classified as positive (35) or negative (33). The explanation for employing this dataset is that reported speech (a person referring to another person or event) represents a direct and unbiased expression of opinions that does not depend on the interpretation of the reader in the majority of cases.

At the present moment, there are some lexicons annotated with affect and polarity at the sense level and are based on WordNet [13]: WordNet-Affect [14], SentiWordNet [1] and Micro-WNOp [15]. WordNet-Affect, an extension of WordNet Domains, is a hierarchy of affective domain labels which were developed by selecting suitable synsets from WordNet which represent affective concepts and dividing them into subsets of affective data. In SentiWordNet, each synset in WordNet has assigned three values of sentiment: positive, negative and objective, whose sum is 1. For example, the synset HAPPY#3 (marked by good fortune; "a felicitous life"; "a happy outcome"), is annotated as Positive = 0.875, Negative = 0.0 and Objective = 0.125. This resource was created through a mix of linguistic techniques and statistical classifiers. It was semi-automatically built so all the results were not manually validated and some resulting classifications can appear incorrect. Finally, the Micro-WNOp corpus is composed by 1105 WordNet synsets manually annotated in a manner that is similar to SentiWordNet.

4.2 Word-based Method Applied to Polarity Classication

For the first approach, as in [12], each of the employed resources were mapped to four categories, which were given different scores - positive (1), high positive (4), negative (-1) and high negative (-4). On the one hand, the approach is motivated by the same mapping done in [12], and, on the other, on the wish to mantain the "idea" of the lexicons employed - that the same term may have different strengths of polarity and that two different terms even if they have the same polarity, may differ in intensity.

The words belonging to the WordNet-Affect categories of anger and disgust were grouped as in [12] under high negative, fear and sadness were considered negative, joy was taken as containing positive words and surprise as highly positive; SentiWordNet and Micro-WNOp contain positive and negative scores between 0 and 1 and in their case, the mapping was done in the following manner: the words that have senses with positive scores lower than or equal to 0.5 to the positive category, the scores higher than 0.5 to the high positive set, the negative scores lower than or equal to 0.5 to the negative category and the ones higher than 0.5 to the high negative set. See Table 1 for the statistics of the categories built from each resource. The last row corresponds to the union of the categories for all resources.

Table 1. Statistics of the categories used by the word-based method.

Resource	**Positive**	**Negative**	**High Positive**	**High Negative**
WN-Affect	192	215	73	201
Micro-WNOp	436	396	409	457
SentiWN	23133	22144	2462	5279
SentiWN+Micro-WNOp+WN-Affect	23394	22442	2804	5713

Finally, the polarity value of each of the quotes was computed as sum of the values of the words identified; a positive score leads to the classification of the quote as positive, whereas a final negative score leads to the system classifying the quote as negative. A quote is classified as neutral if the score is equal to 0. Note that no word sense disambiguation is done in this method, rather a word is incorporated into a category depending on the annotation of its senses. Thus, a same word can be included to several categories and the word-to-sense relationship is lost. The results of this approach are shown in Table 2.

Table 2. Classification results of the method without WSD.

Resources	**P+**	**P-**	**R+**	**R-**	**F1+**	**F1-**	**P**	**R**	**F1**	**Cov**
WN-Affect	0.75	0.60	0.08	0.09	0.15	0.16	0.67	0.09	0.15	0.13
Micro-WNOp	0.57	0.67	0.57	0.18	0.57	0.28	0.59	0.38	0.46	0.65
SentiWN	0.55	0.46	0.48	0.36	0.51	0.41	0.51	0.43	0.46	0.84
SentiWN+WN-Affect	0.55	0.46	0.48	0.36	0.51	0.41	0.51	0.43	0.46	0.84
SentiWN + Micro-WNOp	0.53	0.48	0.54	0.33	0.54	0.39	0.51	0.44	0.47	0.87
All	0.53	0.45	0.54	0.33	0.53	0.38	0.50	0.44	0.46	0.88

4.3 WSD Method Applied to polarity classification

This approach based on word sense disambiguation to determine the polarity of opinions was previously presented in [16], where it was evaluated over the SemEval Task No. 14: Affective Text data, outperforming the results obtained by both unsupervised and supervised systems participating in the competition.

Word Sense Disambiguation (WSD) is an intermediate task of Natural Language Processing. It consists in selecting the appropriate meaning of a word given the context in which it occurs [17].

The approach is based on the assumption that the same word, in different contexts, may not have the same polarity. For example, the word "drug" can be positive, negative or objective, depending on the context where it appears (e.g., "she takes drugs for her heart" (objective), "to be on drugs" (negative)). Bearing in mind this need to appropiately identify the correct sense, we use a word sense disambiguation algorithm to obtain the correct sense of the words in the opinion and subsequently obtain the polarity of the senses from resources based on senses annotated with valence and emotions. The WSD-based method also handles negations and other polarity shifters obtained from the General Inquirer dictionary.

For the disambiguation of the words, we use the method proposed in [18], which relies on clustering as a way of identifying semantically related word senses. In this WSD method, the senses are represented as signatures built from the repository of concepts of WordNet. The disambiguation process starts from a clustering distribution of all possible senses of the ambiguous words by applying the Extended Star clustering algorithm [19]. Such a clustering tries to identify cohesive groups of word senses, which are assumed to represent different meanings for the set of words. Subsequently, clusters that best match the context are selected. If the selected clusters disambiguate all words, the process stops and the senses belonging to the selected clusters are interpreted as the disambiguating ones. Otherwise, the clustering process is performed again (regarding the remaining senses), until a complete disambiguation is achieved.

Once the correct sense for each word on the opinion is obtained, the method determines its polarity regarding the sentiment annotation for this sense in the lexical resource utilized. From SentiWordNet and Micro-WNOp we obtain a positive and a negative value for the target sense (in Micro-WNOp only a part of the synsets are annotated with the polarity, thus the senses that are not annotated are considered to be completely objectives). In the case of WordNet-Affect that is annotated with emotions and not with values of polarity as such, we build a mapping, the senses pertaining to the hierarchy of positive (negative) affective domain labels were assigned a positive value of 1(0) and a negative value of 0(1), respectively.

Finally, the polarity of the opinion is determined from the scores of positive and negative words it contains. To sum up, for each word w and its correct sense s, the positive ($P(w)$) and negative ($N(w)$) scores are calculated as:

$$P(w) = \text{Positive value of } s \text{ in a } \textit{lexical resource}$$

$N(w)$ = Negative value of s in a *lexical resource*

Finally, the global positive and negative scores (S_p, S_n) are calculated as:

$$S_p = \sum_{w:P(w)>N(w)} P(w)$$

$$S_n = \sum_{w:N(w)>P(w)} N(w)$$

If S_p is greater than S_n then the opinion is considered as positive. On the contrary, if S_p is less than S_n the opinion is negative. Finally, if S_p is equal to S_n the opinion is considered as neutral. In Table 3 the results are shown.

Table 3. Classification results of the WSD-based method.

Resources	**P+**	**P-**	**R+**	**R-**	**F1+**	**F1-**	**P**	**R**	**F1**	**Cov**
WN-Affect	1.00	0.75	0.17	0.09	0.29	0.16	0.90	0.13	0.23	0.15
Micro-WNOp	0.58	0.40	0.20	0.06	0.30	0.11	0.53	0.13	0.21	0.25
SentiWN	0.48	0.53	0.46	0.45	0.47	0.49	0.51	0.46	0.48	0.90
SentiWN+WN-Affect	0.50	0.55	0.46	0.48	0.48	0.52	0.52	0.47	0.50	0.90
SentiWN + Micro-WNOp	0.50	0.56	0.49	0.45	0.49	0.50	0.52	0.47	0.50	0.90
All	0.53	0.57	0.51	0.48	0.52	0.52	0.55	0.50	0.52	0.91

5 Discussion

From Table 2, we can observe that the worst results for the word-based approach were obtained using WN-Affect. Note that the categories that are built from this resource contain few words and therefore the coverage of the method is affected (see Table 1 for statistics of the resources). For Micro-WNOp the method improves the coverage but fails in the detection of negative quotes (see low values of R- and F1-). The best results were obtained in the combinations that uses SentiWN; more negative opinions are correctly classified and better coverage is achieved. Combining SentiWN with other resources do not seem to improve the F1 scores, even though the coverage is slightly better. As WN-Affect and Micro-WNOp were built annotating a subset of WordNet senses and SentiWN includes all of these senses, it is likely that the four categories of SentiWN and those built from each combination are not significantly different (see Table 1).

Regarding the approach based on word sense disambiguation, we can observe in Table 3 that for the Micro-WNOp and WN-Affect resources the method obtains very low results, due to the low coverage of the annotated senses; from 115425 synsets in WordNet, only 1105 (0.96%) and 884 (0.77%) are annotated on these resources, respectively. Also, the corpus has 1472 words, of which 1277 are non stop-words and are disambiguated with the senses of WordNet; Micro-WNOp only covers 57 (4.46%) and WN-Affect 18 (1.41%) of these ambiguous words. In spite of the low coverage, the method obtains acceptable precision values when these resources are used. For these reasons, when we use these resources individually, only a few words obtain a value of polarity. On the other hand, the use of SentiWN significantly improves both the F1 scores and the coverage.

Note also, that the combination of several resources obtains better precision, recall and F1 scores. Due to the fact that SentiWN was not manually annotated, some senses are misclassified (e.g., the sense FLU#1 (*an acute febrile highly contagious viral disease*) is annotated as Positive = 0.75, Negative = 0.0 and Objective = 0.25, despite having a lot of negative words in its gloss). These mistakes affect the polarity classification. We suppose that combining WN-Affect and Micro-WNOp with SentiWN reduces this negative influence, and consequently the precision, recall and F1 values are improved. The low coverage of the Micro-WNOp and WN-Affect resources do not allow higher increases in the classification quality.

Finally, Figure 1 shows the comparison of both methods (with and without WSD) for each resource combination with respect to overall F1 measure. As can be seen, the results of the method based on word sense disambiguation are better than those of the bag-of-words approach, except for Micro-WNOp where the WSD-based method is severely affected by the low coverage of this resource. The best F1 score of the WSD-based method is 0.52 and that of the method without WSD is 0.47. Note also that the WSD-based method not only obtains better overall F1, but also a higher coverage (see Tables 2 and 3). This confirms that word sense disambiguation is useful for determining the polarity of a word.

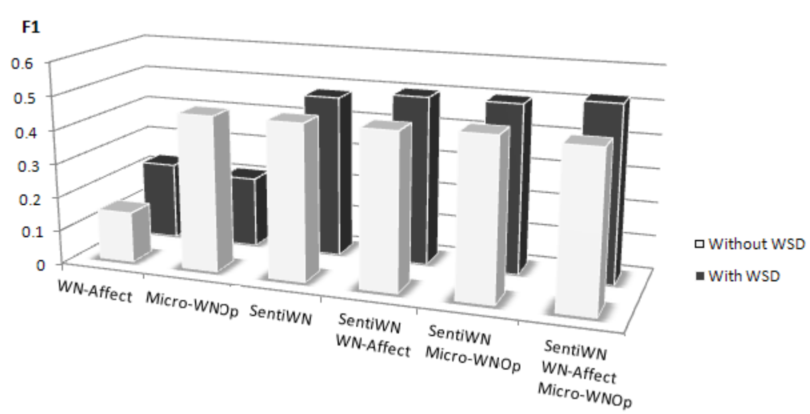

Fig. 1. Overall F1 scores for both methods.

Unlike the WSD-based method, the addition of WN-Affect and Micro-WNOp with respect to use only SentiWN does not contribute to improve the results of the method without WSD. In the first method, this is due to the higher quality in the sense polarity annotation, while in the second one to the absence of differences among the built categories.

From the results of Tables 1 and 2, we can also notice that Micro-WNOp and WN-Affect achieve better overall precision but lower overall recall than SentiWN in both methods. It can be expected due to the manual annotation and the low coverage of these resources, respectively.

Another interesting observation is that the bag-of-words approach leads to better performance when classifying positive quotes, whereas the WSD-based method achieves a good balance between the classification of positive quotes and the negative ones instead. This can be seen in the precision, recall and F1 values for each class.

6 Conclusions and Future Work

In this paper, a comparison between two unsupervised methods for determining the polarity of opinions has been presented. One of them performs the analysis at a word level, whereas the other at a sense level. Studies of the behaviour of both methods were presented over the same corpus and using several public resources for opinion mining. In the experiments, we demonstrate that word sense disambiguation is useful to determine the polarity of opinions.

The use of word sense disambiguation in the polarity classification has pros and cons. The advantages are given in the superiority of the results (with respect to precision, recall and F1) obtained by taking into account the context of the words appearing in the opinion. Also, the polarity detection has the same behavior in both classes, that is, the performance is balanced when positive and negative quotes are classified. Despite that some of the resources have a low coverage, this method obtains a better coverage than the bag-of-words approach.

However, as word sense disambiguation constitutes an intermediate task in the polarity classification, the disambiguation errors could affect the classification quality. This provides further motivation to study in depth this problem, due to the lack of a corpus manually annotated with senses and polarity. Also, the disambiguation algorithm depends on the used knowledge resources and, as we can saw in the experiments, there are no resources that have both high coverage and a good quality in the annotation of sense polarities.

Future work includes the study of alternative methods to extract and classify opinions, working at a syntactic level, or using local contexts, semantic representations of concepts and the modelling of discourse structures. Our idea is to study in a broader context the impact of word sense disambiguation on the performance of opinion mining systems, be it in small texts (such as the ones we have studied in this paper) or larger contexts (on-line discussion forums, blogs or newspaper articles). Another interesting application would be the determination of figurative senses, which are used to express opinions in a more sophisticated manner. To this aim, the application of word sense disambiguation is an essential step.

References

1. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Fifth international conference on Language Resources and Evaluation (LREC 2006). (2006) 417–422
2. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, The Association for Computational Linguistics (1997) 174–181
3. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems **21** (2003) 315–346
4. Kim, S., Hovy, E.: Determining the sentiment of opinions. In: 20th International Conference on Computational Linguistics (ACL 2004), Morristown, NJ, USA, The Association for Computational Linguistics (2004) 1367–1373
5. Takamura, H., Inui, T., Okumura, M.: Extracting emotional polarity of words using spin model. In: 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, US, The Association for Computational Linguistics (2005) 133–140
6. Esuli, A., Sebastiani, F.: Pageranking wordnet synsets: An application to opinion mining. In: 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, The Association for Computational Linguistics (2007) 424–431
7. Wiebe, J., Mihalcea, R.: Word sense and subjectivity. In: 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia, The Association for Computational Linguistics (2006) 1065–1072
8. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **1** (2005) 165–210
9. Akkaya, C., Wiebe, J., Mihalcea, R.: Subjectivity word sense disambiguation. In: Conference on Empirical Methods in Natural Language Processing, Singapore, The Association for Computational Linguistics (2009) 190–199
10. Rentoumi, V., Giannakopoulos, G.: Sentiment analysis of figurative language using a word sense disambiguation approach. In: International Conference on Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria, The Association for Computational Linguistics (2009)
11. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Conference on New Methods in Language Processing. (1994) 44–49
12. Balahur, A., Steinberger, R., v. d. Goot, E., Pouliquen, B., Kabadjov, M.: Opinion mining on newspaper quotations. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference **3** (2009) 523–526
13. Fellbaum, C.: WordNet: an electronic lexical database. MIT Press (1998)
14. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: 4th International Conference on Language Resources and Evaluation, LREC 2004. (2004) 1083–1086
15. Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., Gandini, G.: Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: Language resources and linguistic theory: Typology,

second language acquisition, English linguistics. Franco Angeli Editore, Milano, IT (2007)
16. Martín-Wanton, T., Pons-Porrata, A., Montoyo-Guijarro, A., Balahur, A.: Opinion polarity detection: Using word sense disambiguation to determine the polarity of opinions. In: 2nd International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial Intelligence, Valencia, Spain, INSTICC Press (2010) 483–486
17. Agirre, E., Edmonds, P.: Word Sense Disambiguation: Algorithms and Applications. Volume 33 of Text, Speech and Language Technology. Springer-Verlag New York, Inc. (2006)
18. Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R.: Word sense disambiguation based on word sense clustering. In Coelho, J.S.H., Oliveira, S., eds.: Lecture Notes in Artificial Intelligence, vol. 4140, Springer (2006) 472–481
19. Gil-García, R., Badía-Contelles, J.M., Pons-Porrata, A.: Extended star clustering algorithm. In Sanfeliu, A., Ruiz-Shulcloper, J., eds.: Lecture Notes in Computer Sciences, vol. 2905, 8th Iberoamerican Congress on Pattern Recognition (CIARP), Springer-Verlag (2003) 480–487