Combining semi-supervised and active learning to recognize minority senses in a new corpus

Cristian Cardellino, Milagro Teruel and Laura Alonso i Alemany
Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Argentina

Abstract

In this paper we study the impact of combining active learning with bootstrapping to grow a small annotated corpus from a different, unannotated corpus. The intuition underlying our approach is that bootstrapping includes instances that are closer to the generative centers of the data, while discriminative approaches to active learning include instances that are closer to the decision boundaries of classifiers.

We build an initial model from the original annotated corpus, which is then iteratively enlarged by including both manually annotated examples and automatically labelled examples as training examples for the following iteration. Examples to be annotated are selected in each iteration by applying active learning techniques.

We show that intertwining an active learning component in a bootstrapping approach helps to overcome an initial bias towards a majority class, thus facilitating adaptation of a starting dataset towards the real distribution of a different, unannotated corpus.

1 Introduction and Motivation

Verbal sense disambiguation (VSD) is a crucial task for deep language processing tasks, specially those that could benefit from information provided by subcategorization frames, like machine translation, question answering or information extraction. In recent years VSD has achieved major improvements in performance, at least for English, where big annotated corpora are available and with thorough studies on characterization of examples [Chen and Palmer, 2009; Croce *et al.*, 2012; Kawahara and Palmer, 2014].

However, the distribution of senses is imbalanced, for verb senses as it is for all natural language phenomena, following Zipf's law [Zipf, 1949]. This implies that many verbs or verbal senses occur infrequently, and therefore examples for such cases are few in annotated corpora. Semi-supervised approaches like bootstrapping are useful to grow a small training corpus, but have a strong bias towards what they already know, thus not including examples of rare cases that were not included in the starting corpus, or even discarding minority distinctions if better accuracy figures can be obtained by assigning all examples to a majority class. This is even more acute when we try to adapt a starting model to a different corpus: most likely, only examples from the majority class will be found.

To prevent bootstrapping methods from falling into this bias and recognizing the actual distribution of senses in the new corpus, some sort of guiding mechanism has to be introduced. What we propose in this paper is to resort to active learning techniques as a complement to bootstrapping to recognize new and minority senses, and help grow minority classes. Active learning has shown to be useful in verbal sense disambiguation for English [Chen *et al.*, 2006].

The rest of the paper is organized as follows. In the next section we describe some relevant work that addresses the problem of VSD and WSD using semi-supervised methods. Then, in section 3 we detail the iterative bootstrapping algorithm we use on our set of experiments, and how it is complemented by active learning techniques. In section 4 we present the experiments and results we achieve in both bootstrap alone and combined with active learning. Finally, we conclude with some remarks and our lines of current and future work.

2 Relevant Work

This work builds on two main areas of previous work: bootstrapping techniques and active learning as applied to word sense disambiguation and verbal sense disambiguation. We analyze them in what follows.

The landmark work on bootstrapping for word sense disambiguation is the 1995 Yarowsky paper [Yarowsky, 1995]. In his work, Yarowsky builds a disambiguation model based on the words co-occurring with manually labeled examples. Then, this model is applied to unlabeled examples. Examples that can be assigned a sense by the model are then incorporated as training examples, and a new model is trained. This process is iteratively applied until a termination condition is reached, namely, no new examples can be assigned a sense or the reliability of the evidence found by the model is too low. After each iteration, the resulting model has arguably bigger coverage than previous versions. Therefore, this method is useful to build a real-life tool out of a limited number of examples.

Ye and Baldwin [Ye and Baldwin, 2006], use SPs extracted with a Semantic Role Labeler (SRL) for VSD. Their VSD framework is based upon three components: extraction of disambiguating features, selection of the best disambiguating feature with respect to unknown data and the tuning of the machine learner's parameters. For their study they use a Maximum Entropy algorithm [Berger *et al.*, 1996].

Another work on English VSD is the one by [Chen and Palmer, 2009], presenting a high-performance broad-coverage supervised word sense disambiguation (WSD) system for English verbs that uses linguistically motivated features and a smoothed maximum entropy machine learning model. [Kawahara and Palmer, 2014] presented a supervised method for verb sense disambiguation based on VerbNet. Contrary to the most common VSD methods, which create a classifier for each verb that reaches a frequency threshold, they created a single classifier to be applied to rare or unseen verbs in a new text. Their classifier also exploits generalized semantic features of a verb and its modifiers in order to better deal with rare or unseen verbs.

[Chen *et al.*, 2006] show that Active Learning for verbal sense disambiguation for English is useful, sometimes reducing by half the amount of examples needed to achieve a certain accuracy, but tends to suffer from overfitting. They propose to deal with overfitting by dimensionality reduction via feature extraction. In our approach, we expect that the sheer amount of examples introduced by bootstrapping will help reduce overfitting.

[Dligach and Palmer, 2011] explore the benefits of using an unsupervised language model to select rare examples in cases of a skewed distribution of classes, with successful results. This language model is acting as a density estimation method to locate generative sources of examples, thus overcoming the bias of discriminative methods towards majority classes in skewed distributions.

In our approach we explore the combination of bootstrap as a density estimation method and a classical discriminative method, *uncertainty sampling*, to overcome the majority class bias while keeping a good learning rate.

3 Combining bootstrapping and active learning

The intuition underlying our approach is that bootstrapping and active learning can complement each other. Bootstrapping builds upon *certainties* of the classifier, thus incorporates examples that are similar to already known examples. Conversely, active learning, or at least *discriminative* approaches to active learning (as opposed to density estimation approaches like [Xu *et al.*, 2003]) tend to incorporate examples that are most dissimilar to already known examples. In other words, bootstrapping finds instances that are closer to the generative centers of the data, while active learning finds instances that are closer to the decision boundary of the classifier.

In our approach we combine classical bootstrapping with an active learning method called *uncertainty sampling*. Uncertainty sampling is one of many pool-based approaches that rank unannotated examples by their expected value if annotated by a human oracle. In uncertainty sampling, examples where the automated classifier is most uncertain are ranked higher. Then, the n top ranking examples are provided for a human judge to annotate.

Our algorithm starts from an initial annotated corpus, 20% of which is reserved for testing and the remaining 80% is used as a starting training corpus. We learn a classification model from this training set and apply it to a large unannotated corpus. The classifier predicts a class with certain confidence for each instance. If the confidence of the classifier is over a given threshold, that instance is automatically added to the annotated corpus as a new example, with the predicted class as the label. This is the bootstrapping approach to growing the annotated corpus.

Instances are also ranked by inverse certainty order, and the n instances with least certainty are provided to a human judge to annotate. This is the uncertainty sampling, active learning approach to growing the annotated corpus.

With the new annotated corpus (the annotated corpus from the last iteration plus the new instances of this iteration), we use the information and get new features to train a new model, starting a new iteration of the method.

Before running each iteration, we evaluate the new training model using 10-fold cross-validation on the training set accumulated to that moment. If the accuracy obtained by cross-validation is below a certain threshold, iterations are terminated. Another stopping criterion is when the algorithm reaches a user-defined number of iterations. If no stopping criterion is met, we use the new model for the next iteration.

When all the iterations have finished we classify the test instances with the model from the last model to asses the impact of the initial model against the last model in the test corpus.

4 Experiments

We used a SVM classifier with a feature selection preprocess, which, as pointed out by [Chen and Palmer, 2009], improves the performance of the classifier by eliminating the tendency to overfitting. As a baseline we classified all the instances are classified as the most frequent sense. This baseline is specially interesting because one of the drawbacks of having bootstrap as an automatic approach is the bias towards the most frequent class which is common place in skewed distribution. In our corpus, in average, the most frequent sense of each verb took about 65% of the examples before bootstrap was applied, and 75% afterwards.

The annotated corpus for the initial model is SenSem [Alonso *et al.*, 2007], a corpus of verb senses for Spanish, with manually annotated examples for the 250 most frequent verbs of Spanish. We discard those lemmas with a single sense, roughly 10% of the lemmas. The corpus is preprocessed using Freeling [Padró and Stanilovsky, 2012] to obtain information on chunks, syntactic functions and dependency triples that makes the annotated examples comparable to unannotated examples. The unannotated corpus is the Wikicorpus [Reese *et al.*, 2010], also preprocessed with Freeling.

In each iteration of the VSD Algorithm, we use the previous model to classify a randomly selected sample of instances

Verb	Accuracy		Precision		Recall	
	Before	After	Before	After	Before	After
abrir	55.00	80.00	0.54	0.84	0.55	0.80
acercar	61.90	90.48	0.59	0.93	0.62	0.91
dedicar	56.00	88.00	0.56	0.89	0.56	0.88
detener	76.00	96.00	0.72	0.97	0.76	0.96
encontrar	73.91	100.00	0.76	1.00	0.74	1.00
hallar	52.17	86.96	0.52	0.91	0.52	0.87
interpretar	44.00	80.00	0.42	0.85	0.44	0.80
ir	63.64	81.82	0.57	0.85	0.64	0.82
limitar	82.61	100.00	0.81	1.00	0.83	1.00
seguir	57.14	85.71	0.59	0.89	0.57	0.86

Table 1: Classifier results before and after bootstrapping task

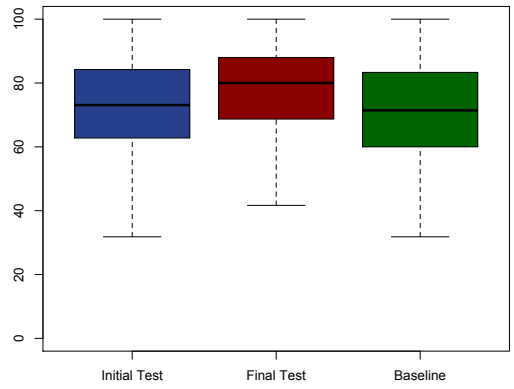

Figure 1: Mean of the accuracies obtained for all lemmas before (initial) and after (final) the 10 bootstrapping iterations, and the accuracy of the most frequent sense baseline.

from the unannotated corpus. We had to use a sample rather than the whole corpus for reasons of hardware limitations. However, we feel the random sampling reflects the behavior of the totality of the unannotated corpus, since this instances are randomly selected in each iteration.

4.1 Features

We discarded information about semantic roles and constituents that was provided by the annotated corpus, because we cannot provide such information for the unannotated corpus, and annotated and unannotated examples would be incomparable. We then choose to parse both corpuses with Freeling to use chunks, syntactic functions, and dependency triples Freeling returned, besides the bags-of-ngrams both corpus already provided; this way we were able to get a richer set of features than just bags-of-ngrams. The selected features for the final experiments were: bags-of-ngrams (unigrams, bigrams and trigrams), chunks, syntactic functions, and dependency triples. Features were also filtered so only those appearing a minimum of 5 times in the annotated corpus were used for the experiments. This provides a reduction of dimensionality that makes the problem more tractable and less prone to overfitting.

4.2 Bootstrapping approach

For the bootstrapping task we set a threshold confidence of 90% to automatically add a new example. The threshold accuracy to stop the iterations after the 10-fold cross-validation was set on 50% and the maximum amount of iterations was set to 10.

Figure 1 shows the average accuracy of the classifier before and after applying bootstrap, without an active learning component, together with the baseline. As we can observe from the Figure, the performance is improved importantly, above the most frequent class baseline.

Table 1 shows the detail of the accuracy for some of the verbs, before and after the bootstrap iterations. As we can see, these verbs achieve quite better results taking advantage of new automatically added examples.

4.3 Bootstrapping with active learning

When combined with active learning, in each iteration we included not only automatically selected examples (the ones where the classifier showed more certainty) but also manually annotated examples. In each iteration, the human judge was provided with the 10 examples with lowest certainty to annotate. This manual annotation was carried out for only 4 of the verb lemmas in SenSem: *apuntar* (*point out*, *note down*), *creer* (*believe*), *escuchar* (*listen*, *hear*) and *hablar* (*speak*).

In Figure 2 we can compare the performance of bootstrapping alone and bootstrapping with the active learning component. From the plot in the left we can see that the accuracy of bootstrapping alone is overall above the accuracy of boostrapping with active learning added. This results seem to indicate that active learning is actually introducing a decrease in performance by comparison to bootstrap alone. However, we can also see that root mean squared error is reduced if active learning is added to bootstrapping. We can understand this with the more detailed error analysis that follows.

In Figure 3 we can see the detail of the f-scores obtained for three verbs for which we carried out manual annotation with active learning. The fourth verb is not displayed because only one of its senses was represented in the annotated examples. F-scores are displayed to represent accuracies by class, in this case, by sense. In these plots we can see that when active learning was included, the classifier was able to distinguish more senses than those in the starting corpus, thus effectively adapting to the new, unannotated corpus. In some case, senses that were distinguished by the bootstrap only approach showed a decrease in f-measure in the boostrap-and-active-learning approach, but this decrease was not dramatic.

A detail of confusion matrices before and after the iterations with bootstrap and active learning can be seen in Figure 4. In these matrices we can see that initially the classifier classified most of the examples in the majority class, achieving an accuracy of 68%. After the 10 iterations with bootstrap and uncertainty sampling, the performance dropped more than 10 points, to 56%, but the bias towards the majority class was gone, and instances were classified more evenly

across classes, with many instances classified in the classes where they actually belonged. So, even if accuracy decreased, we obtained a qualitative improvement in performance and successfully recognized classes that were actually present in the new corpus, and not in the starting corpus.

5 Conclusions and Future Work

In this paper we present an initial approach to combine bootsrapping and active learning to adapt a verbal sense disambiguation classifier from a small annotated corpus to a bigger, unannotated corpus. This combination of techniques is capable of recognizing minority senses in the new corpus, even if it implies losing some overall accuracy.

Future work includes experiments where examples will be characterized by richer features, including semantic roles and synsets of verb constituents, in the line of [Kawahara and Palmer, 2014], [Ye and Baldwin, 2006] and [Chen and Palmer, 2009].

References

[Alonso *et al.*, 2007] L. Alonso, J.A. Capilla, I. Castellón, A. Fernández, and G. Vázquez. The sensem project: Syntactico-semantic annotation of sentences in spanish. In N. Nicolov et al., editor, *Selected papers from RANLP 2005*, pages 89–98. John Benjamins, 2007.

[Berger *et al.*, 1996] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March 1996.

[Chen and Palmer, 2009] Jinying Chen and Martha S Palmer. Improving english verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Language Resources and Evaluation*, 43(2):181–208, 2009.

[Chen *et al.*, 2006] Jinying Chen, Andrew Schein, Lyle Ungar, and Martha Palmer. An empirical study of the behavior of active learning for word sense disambiguation. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 120–127, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Croce *et al.*, 2012] Danilo Croce, Alessandro Moschitti, Roberto Basili, and Martha Palmer. Verb classification using distributional similarity in syntactic and semantic structures. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 263–272, 2012.

[Dligach and Palmer, 2011] Dmitriy Dligach and Martha Palmer. Good seed makes a good crop: accelerating active learning using language modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 6–10. Association for Computational Linguistics, 2011.

[Kawahara and Palmer, 2014] Daisuke Kawahara and Martha Palmer. Single classifier approach for verb sense disambiguation based on generalized features. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[Padró and Stanilovsky, 2012] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[Reese *et al.*, 2010] Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[Xu *et al.*, 2003] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. Representative sampling for text classification using support vector machines. In *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR 2003)*, volume 2633 of *LNCS*, pages 393–407, Pisa, Italy, April 2003. Springer.

[Yarowsky, 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL-95*, pages 189–196, Cambridge, MA, 1995. ACL.

[Ye and Baldwin, 2006] Patrick Ye and Timothy Baldwin. Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 139–148, Sydney, Australia, November 2006.

[Zipf, 1949] George Zipf. *Human Behavior and the Principle of Least Effort*. Addison–Wesley, Cambridge, MA, 1949.

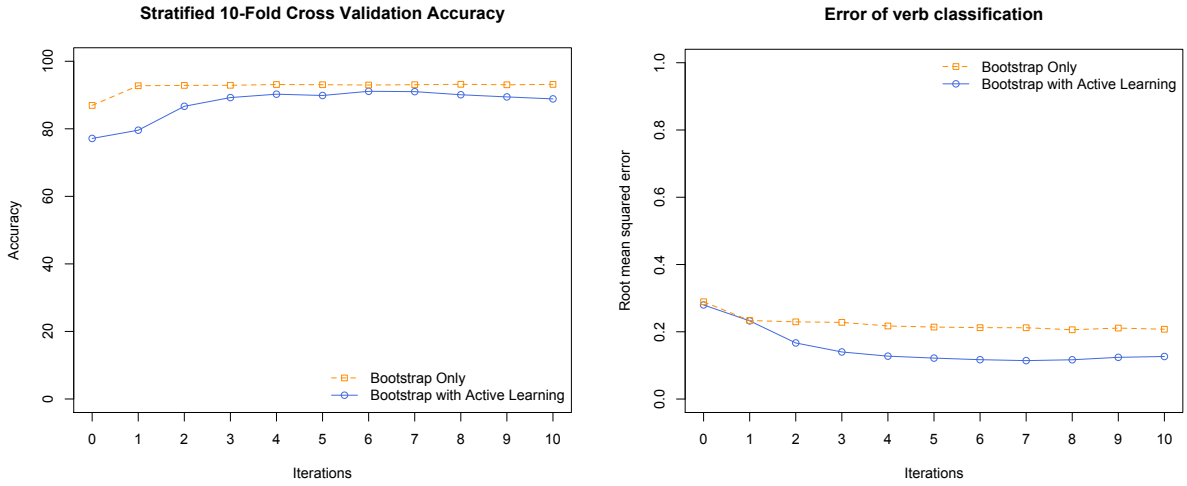

Figure 2: Mean accuracy (left) and error (right) of different iterations of bootstraping with (circles) and without (squares) active learning.

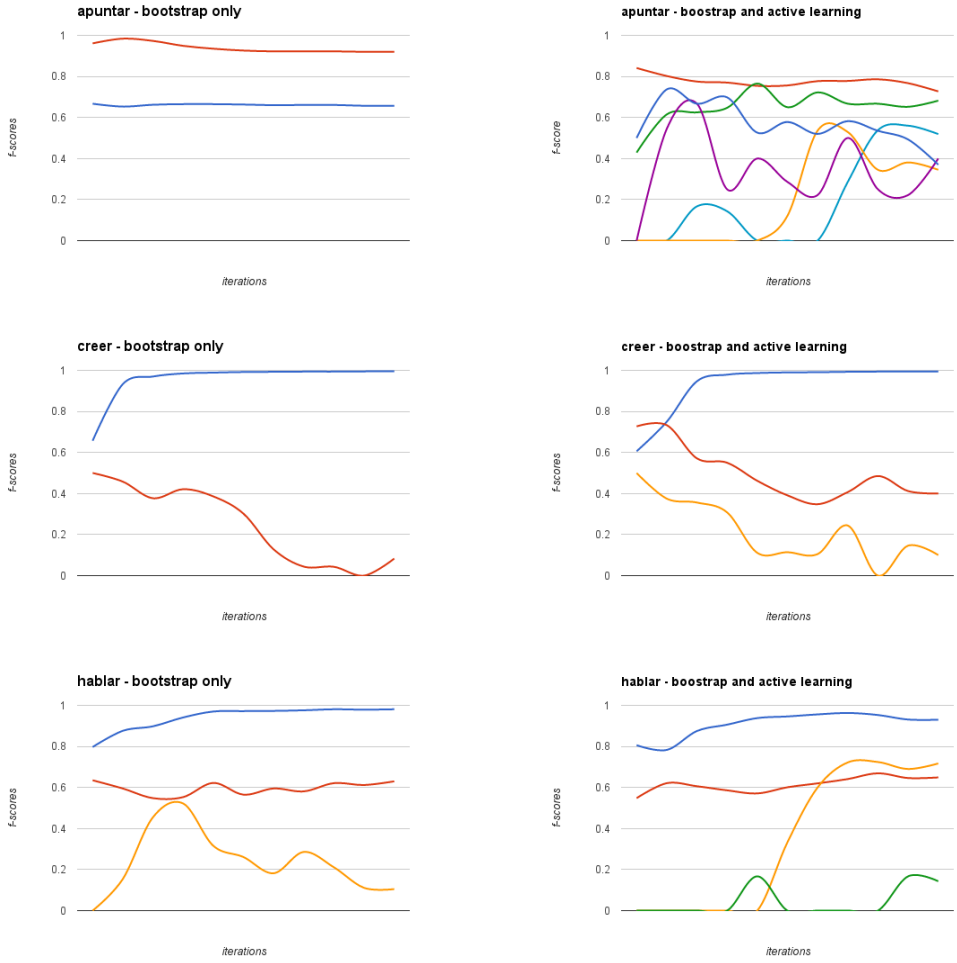

Figure 3: F-score obtained for each of the senses for ambiguous lemmas, across iterations, with and without active learning.

a	b	c	d	e	f	g	h	i	j	<-- classified as
0	0	0	0	6	0	1	0	0	0	| a = apuntar-1
0	4	0	0	6	0	0	0	0	0	| b = apuntar-10
0	0	0	0	4	0	0	0	0	0	| c = apuntar-2
0	0	0	0	4	0	0	0	0	0	| d = apuntar-3
0	1	0	0	61	0	0	0	0	0	| e = apuntar-4
0	0	0	0	0	0	1	0	0	0	| f = apuntar-5
0	1	0	0	1	0	3	0	1	0	| g = apuntar-6
0	0	0	0	0	0	0	0	0	0	| h = apuntar-7
0	0	0	0	0	0	3	0	0	0	| i = apuntar-8
0	0	0	0	1	0	0	0	1	0	| j = apuntar-9

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
7	2	0	1	0	9	0	0	0	0	0	| a = apuntar-1
0	15	0	9	0	13	0	0	0	0	0	| b = apuntar-10
0	3	0	2	0	0	0	0	0	0	0	| c = apuntar-11
0	10	0	10	0	10	0	2	0	0	0	| d = apuntar-2
0	0	0	1	0	6	0	0	0	0	0	| e = apuntar-3
0	12	0	3	0	84	0	3	0	0	0	| f = apuntar-4
0	0	0	0	0	1	0	1	0	1	0	| g = apuntar-5
0	2	0	0	0	2	0	15	0	1	0	| h = apuntar-6
0	0	0	0	0	1	0	0	0	0	0	| i = apuntar-7
0	0	0	0	0	2	0	2	0	2	0	| j = apuntar-8
1	0	0	0	0	1	0	1	0	0	0	| k = apuntar-9

Figure 4: Confusion matrices with the results of classification of the test corpus in the starting and final iterations of the bootstrapping and active learning approach for the verb "*apuntar*", showing the classification of instances across different senses in the last iteration, in contrast with the classification of most instances in the majority sense in the first iteration.