# 19 Comparison of Tagging Strategies

## 19.1 Idea and Setup

As explained in Chapter 10, *tagging strategies* are a core constituent of classification-based IE approaches, necessary to translate between classes assigned to individual tokens and attribute values that might span multiple tokens. Except for the *Triv* (Trivial) strategy, all the tagging strategies introduced in Sec. 10.2 are able to handle this translation correctly in all situations. But this does not mean that the results reached by using one of them will stay the same when another strategy is employed instead. Strategies differ in the way they partition tokens into class labels, and these differing distributions of class labels may make the problem harder or easier for the used classification algorithm (which can only operate on class labels, without having any knowledge of the underlying attribute values they represent).

So far, the differences in extraction results this causes have never been systematically investigated, to our knowledge. Other classification-based approaches always use a single specific strategy (cf. Sec. 4.4)—we may suppose that some of the authors made some tests to choose among strategies, but there are no reports of results.

In this chapter, we will compare the different strategies to find out whether and how ofter there are significant differences in the results reached by using different strategies in an otherwise identical setup. We also want to find out whether our own choice of using *IOB2* tagging as the default strategy (as for the results reported in the last chapter) is reasonable or whether it would make more sense to use another strategy as default.

To do the comparison, we have used the same corpora and same setup as in the last chapter. Except for varying the tagging strategy, all system settings are identical in all tests. For significance testing, we have applied a paired two-tailed Student's T-test on the F-measure results, without assuming the variance of the two samples to be equal.

## 19.2 Comparison Results

Tables 19.1 and 19.2 list the F-measure results (in percent) reached for both corpora using incremental (online) and batch (iterative) training. It can be seen that batch training generally leads to an improvement compared to incremental training, but in many cases the improvement is small. For the *Corporate Acquisitions* corpus, the batch results of the best strategies are better than any other published results we are aware of; for the *CMU Seminar Announcements*, they are only beaten by the *ELIE* system [Fin04a, Fin04b].

| **Strategy** | **IOB2** | **IOB1** | **Triv** | **BIE** | **BIA** | **BE** |
|---|---|---|---|---|---|---|
| **Seminar Announcements** | | | | | | |
| etime | 96.3 | 93.2 | 93.3 | 92.7 | **97.1** | 92.6 |
| location | 80.1 | 77.7 | 77.3 | 75.7 | **80.7** | 77.1 |
| speaker | **81.0** | 74.6 | 74.1 | 79.2 | **81.0** | 80.3 |
| stime | **99.3** | 98.3 | 98.2 | 98.9 | **99.3** | 98.6 |
| **Corporate Acquisitions** | | | | | | |
| acqabr | 51.7 | 51.3 | **52.0** | 44.8 | 51.8 | 46.2 |
| acqloc | **27.3** | 22.3 | 21.8 | 15.5 | 26.6 | 13.1 |
| acquired | 49.2 | 49.5 | **49.9** | 48.9 | 49.2 | 49.4 |
| dlramt | 60.9 | 60.0 | 60.0 | 59.6 | 60.6 | **62.8** |
| purchabr | 55.3 | 54.0 | 54.2 | 46.1 | **55.8** | 50.4 |
| purchaser | **51.6** | 49.5 | 49.8 | 47.8 | 51.5 | 50.7 |
| seller | 26.0 | 30.5 | **31.1** | 24.4 | 25.7 | 24.0 |
| sellerabr | 24.0 | **29.5** | 28.8 | 14.9 | 24.0 | 20.5 |
| status | 53.0 | 50.1 | 50.0 | 50.9 | **53.5** | 51.2 |

Table 19.1: F-measure Percentages for Incremental Training

| **Strategy** | **IOB2** | **IOB1** | **Triv** | **BIE** | **BIA** | **BE** |
|---|---|---|---|---|---|---|
| **Seminar Announcements** | | | | | | |
| etime | 97.1 | 92.4 | 92.0 | 94.4 | **97.3** | 93.6 |
| location | 81.7 | **81.9** | 81.6 | 77.8 | **81.9** | 82.3 |
| speaker | 85.4 | 82.0 | 82.0 | 84.2 | **86.1** | 83.7 |
| stime | **99.3** | 97.9 | 97.7 | 98.6 | **99.3** | 99.0 |
| **Corporate Acquisitions** | | | | | | |
| acqabr | 55.0 | 53.8 | 53.9 | 48.3 | **55.2** | 50.2 |
| acqloc | 27.4 | **29.3** | **29.3** | 15.7 | 27.4 | 18.0 |
| acquired | 53.5 | **55.7** | 55.5 | 54.8 | 53.6 | 53.7 |
| dlramt | 71.7 | 71.5 | **71.9** | 71.0 | 71.7 | 70.5 |
| purchabr | **58.1** | 56.1 | 57.0 | 47.3 | 58.0 | 51.8 |
| purchaser | 55.7 | 55.3 | **56.2** | 52.7 | 55.7 | 55.5 |
| seller | 31.8 | 32.7 | **34.7** | 27.3 | 30.1 | 32.5 |
| sellerabr | 25.8 | 28.0 | **28.9** | 16.8 | 24.4 | 21.4 |
| status | 56.9 | **57.4** | 56.8 | 56.1 | **57.4** | 55.2 |

Table 19.2: F-measure Percentages for Batch Training

| Strategy | IOB1 | Triv | BIE | BIA | BE |
|---|---|---|---|---|---|
| etime | o (81.6%, –) | o (85.3%, –) | − (98.4%, –) | o (68.6%, +) | o (90.6%, –) |
| location | o (84.3%, –) | o (90.5%, –) | − (98.9%, –) | o (55.8%, +) | − (98.7%, –) |
| speaker | − (98.1%, –) | − (95.3%, –) | o (46.7%, –) | o (1.4%, –) | o (20.8%, –) |
| stime | o (92.9%, –) | − (96.9%, –) | o (75.9%, –) | o (0.0%, =) | o (85.4%, –) |
| acqabr | o (19.8%, –) | o (12.7%, +) | − (98.8%, –) | o (2.2%, +) | − (99.4%, –) |
| acqloc | o (75.0%, –) | o (77.8%, –) | − (98.1%, –) | o (11.2%, –) | − (99.3%, –) |
| acquired | o (17.7%, +) | o (33.6%, +) | o (9.0%, –) | o (0.3%, –) | o (8.9%, +) |
| dlramt | o (6.6%, –) | o (6.5%, –) | o (5.3%, –) | o (2.9%, –) | o (15.1%, +) |
| purchabr | o (45.1%, –) | o (37.8%, –) | − (99.9%, –) | o (14.7%, +) | o (94.0%, –) |
| purchaser | o (62.1%, –) | o (54.8%, –) | o (87.3%, –) | o (6.6%, –) | o (33.8%, –) |
| seller | o (64.3%, +) | o (72.1%, +) | o (20.1%, –) | o (2.8%, –) | o (24.6%, –) |
| sellerabr | o (68.0%, +) | o (64.9%, +) | o (91.9%, –) | o (0.8%, –) | o (45.2%, –) |
| status | o (68.8%, –) | o (70.7%, –) | o (71.7%, –) | o (18.5%, +) | o (64.7%, –) |

Table 19.3: Incremental Training: Significance of Changes Compared to *IOB2*

Tables 19.3 and 19.4 analyze the performance of each tagging strategy for both training regimens, using the popular *IOB2* strategy (our default strategy) as a baseline. The first item in each cell indicates whether the strategy performs significantly better ("+") or worse ("–") than *IOB2* or whether the performance difference is not significant at the 95% level ("o"). In brackets, we show the significance of the comparison and whether the results are better or worse than *IOB2* when significance is ignored.

Considering these results, we see that the *IOB2* and *BIA* strategies are best. No strategy is able to significantly beat the *IOB2* strategy on any attribute, neither with incremental nor batch training. The newly introduced *BIA* (Begin/After) strategy is the only one that is able to compete with *IOB2* on all attributes.

The *IOB1* and *Triv* strategies come close, being significantly worse than *IOB2* only for one or two attributes. The two-classifier *BE* (Begin/End) strategy is weaker, being significantly outperformed on three (incremental) or four (batch) attributes. Worst results are reached by the *BIE* strategy, where the difference is significant in about half of all cases. We suppose that the bad performance might be caused by the fact that *BIE* requires $4n + 1$ classes (where $n$ is the number of attributes), more than any other strategy. The increased complexity of using many similar classes might "confuse" the classifier by introducing subtle and hard to detect differences.

The good performance of *BIA* is interesting, since this strategy is new and has never been used before (to our knowledge). The *Triv* (Trivial) strategy would have supposed to be weaker, considering how simple this strategy is.

## 19.3 Analysis

Our results indicate that the choice of a tagging strategy, while not crucial, should not be neglected when implementing a statistical IE system. The *IOB2* strategy,

| Strategy | IOB1 | Triv | BIE | BIA | BE |
|---|---|---|---|---|---|
| etime | o (87.3%, –) | o (91.8%, –) | o (95.0%, –) | o (18.5%, +) | – (96.9%, –) |
| location | o (18.8%, +) | o (0.5%, –) | – (98.9%, –) | o (22.4%, +) | o (50.3%, +) |
| speaker | – (98.0%, –) | – (99.1%, –) | o (67.0%, –) | o (55.2%, +) | o (88.8%, –) |
| stime | o (82.9%, –) | o (84.4%, –) | o (82.2%, –) | o (11.5%, –) | o (73.4%, –) |
| acqabr | o (49.7%, –) | o (45.8%, –) | – (99.7%, –) | o (6.8%, +) | – (97.9%, –) |
| acqloc | o (56.3%, +) | o (54.0%, +) | – (99.9%, –) | o (1.1%, +) | – (99.4%, –) |
| acquired | o (91.5%, +) | o (84.8%, +) | o (67.9%, +) | o (3.5%, +) | o (8.4%, +) |
| dlramt | o (5.7%, –) | o (14.3%, +) | o (30.2%, –) | o (3.3%, +) | o (46.9%, –) |
| purchabr | o (77.1%, –) | o (44.0%, –) | – (100.0%, –) | o (6.6%, –) | – (99.5%, –) |
| purchaser | o (24.1%, –) | o (26.3%, +) | – (96.0%, –) | o (2.5%, –) | o (17.5%, –) |
| seller | o (34.8%, +) | o (83.5%, +) | – (96.2%, –) | o (59.2%, –) | o (36.1%, +) |
| sellerabr | o (66.7%, +) | o (76.1%, +) | – (99.7%, –) | o (40.7%, –) | o (90.7%, –) |
| status | o (26.3%, +) | o (1.5%, –) | o (43.2%, –) | o (28.0%, +) | o (76.0%, –) |

Table 19.4: Batch Training: Significance of Changes Compared to *IOB2*

which is very popular, having been used in public challenges such as those of *CoNLL* (Conference on Computational Natural Language Learning) [TKS03] and *JNLPBA* (International Joint Workshop on Natural Language Processing in Biomedicine and its Applications) [Kim04], has been found to be indeed the best of all established tagging strategies. It is rivaled by the new *BIA* strategy which we have introduced as a possible alternative. In typical situations, using one of those strategies should be a good choice—since *BIA* requires more classes, it makes sense to prefer *IOB2* when in doubt. Hence, our choice to use *IOB2* as default strategy is indeed reasonable.

Considering that it is not much worse, the *Triv* (Trivial) strategy which requires only a single class per attribute might be useful in situations where the number of available classes is limited or the space or time overhead of additional classes is high. Logically, this strategy is not equivalent to the other ones, since is cannot always translate correctly between state sequences and label sequences, but in practice this weakness has little effect.

The two-classifier *BE* (Begin/End) strategy is still interesting if used as part of a more refined approach, as done by the *ELIE* system (cf. Sec. 4.4.2 for a more detailed discussion of that approach).