21 Mistake Analysis

While quantitative evaluation with the usual performance metrics such as precision and recall allows comparing different approaches and different variations of an approach, it allows little or no insight into *what* mistakes occur and *why* they occur. Yet these questions are essential for a better understanding of where and how we can expect further improvements in information extraction quality to be made, and which limits might exists for information extraction systems in general.

Looking at comparisons of different IE systems on various tasks (such as the comparisons given in Chap. 17), it appears that extraction quality often depends more on the nature of the attributes to extract and the corpora they are in than on the used system. For most attributes, the results reached by various modern systems are fairly similar, while results for different attributes vary enormously, ranging from F-measures $> 99\%$ for the STIME (start time) attribute in the Seminar corpus down to $\approx 25\%$ for the SELLERABR (seller abbreviation) attribute in the Acquisitions corpus. There is little reason to believe that such differences will disappear and attributes such as the latter will ever reach values $> 99\%$ such as the former; but to understand the reasons for such differences, we need to learn more about the nature and the (likely) causes of the mistakes that occur.

As a step in this direction, this chapter provides an analysis of the mistakes our system made on the the two corpora we have used before. The analysis has been performed on the results reached by batch (iterative) training in the standard setup.

21.1 Mistake Types

The mistakes that might occur in the extraction process can be grouped as follows:

- **Boundary mistakes:** predictions can begin or end earlier or later than the corresponding answer key. In this case, the expected answer is partially extracted, but there are spurious tokens at the begin (*early start*) or end (*late end*) of a prediction, or the first (*late start*) or last (*early end*) tokens of the expected answer are missing. For the evaluation metrics, such boundary errors are counted as full errors, even though a partially correct prediction can still contain useful information.
- **Wrong type**: the algorithm predicts an attribute value of the wrong type, for example by wrongly considering the selling party (SELLER) in a corporate acquisition to be the purchasing party (PURCHASER). Such type confusions can occur in combination with boundary mistakes if there is a partial overlap between an answer key of one type and a prediction of another type.

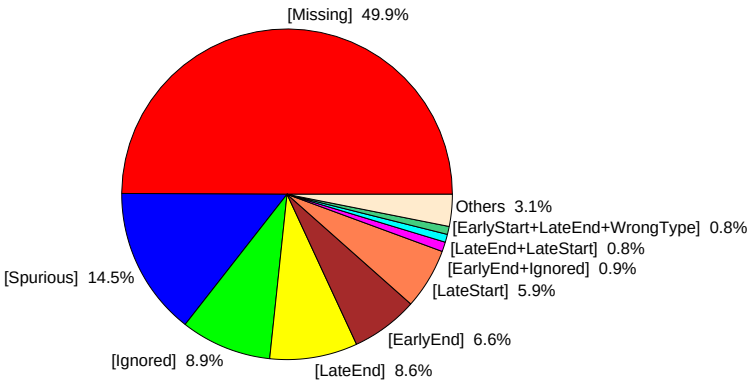


Figure 21.1: Seminar Corpus: Mistakes Combinations

– **Ignored**: For some corpora, including the Seminar and the Acquisitions corpus used in this thesis, only a single answer per attribute and document is expected ("one answer per attribute" evaluation, cf. Sec. 15.2). If there are several prediction candidates, the algorithm has to select one for evaluation, ignoring the others. In this setting it is possible that the selected prediction is wrong but that an *ignored* prediction would have been correct. We have also considered cases where there is an overlap between such an ignored prediction and an answer key of the same or of a different type, i.e., where ignored predictions occur in combination with boundary mistakes or type confusions. We have only considered ignored predictions when the chosen prediction for the same attribute is indeed wrong—otherwise, after all, no mistake occurred.
– **Completely missing** answer keys and **completely spurious** predictions: mistakes where none of the other mistake types applies, i.e., there is no overlap and no type confusion. These mistake types are simply rendered as *missing* and *spurious* in the following charts.

21.2 Distribution of Mistakes

Figure 21.1 shows the mistake combinations that occur in the Seminar Announcements corpus—mistake combinations that occur on average less than once per test run have been combined as *Others*. As stated above, all mistake types except *completely missing* answer keys and *completely spurious* predictions can occur in combination—this results in a high number of possible combination which makes the impact of each mistake type harder to judge. To address this, Fig. 21.2 shows the distribution of mistake types independently of combinations (counting all involved mistake types separately for each combination).

We see that *completely missing* answer keys are by far the largest problem, responsible for almost half of the mistakes. The inverse problem, predictions that are *completely spurious* (no overlap with any answer keys), is the second largest problem. Yet this mistake type is far less frequent, covering less than 15% of all mistakes. Generally, our algorithm tends to favor precision over recall and is more likely to ignore

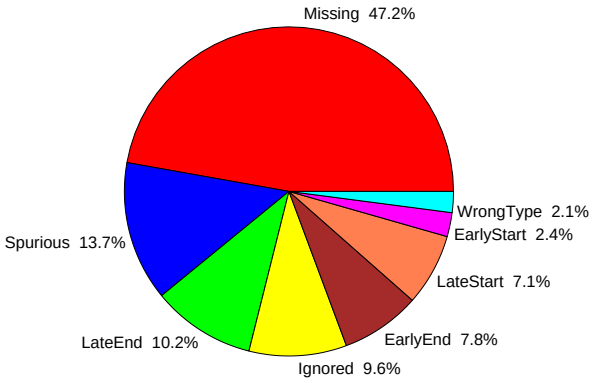


Figure 21.2: Seminar Corpus: Distribution of Mistake Types

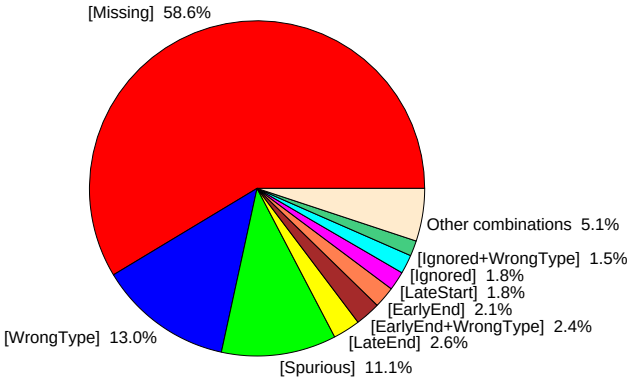
than to extract dubious text fragments—the discrepancy between completely missing and completely spurious attribute values confirms this.

About 9% of the mistakes are predictions that would have been correct, had they not been *ignored* by the extraction algorithm in favor of a more likely (but wrong) alternative. This indicates that the probability estimates assigned by the extraction algorithm should not be neglected when "one answer per attribute" evaluation is used (i.e., the algorithm must choose one among all possible candidates)—if we found a way to improve the estimates we could reduce the frequency of this mistake.

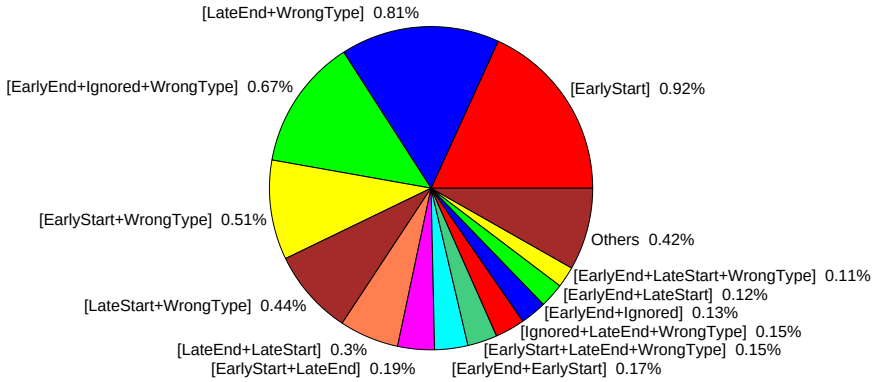
The remaining mistakes have almost all to do with misplaced borders—26% altogether. Correctly detecting the end of attribute values seems to be more difficult than detecting the start: both *late end* (additional trailing tokens in a prediction) and *early end* (prediction missed last tokens) mistakes are more frequent than *late start* mistakes (where leading tokens are missing); *early start* mistakes (leading garbage) are rare (2.4% of all) and usually occur in combination with other mistakes. We will later explore in more detail when and why such boundary mistakes as well as other mistakes occur (Sec. 21.4).

There is one last type of mistakes which is almost invisible in Fig. 21.1: in the Seminar corpus, only a few *wrong type* mistakes occur—about 2% of all mistake types (cf. Fig. 21.2), most of which occur in combination with other mistake types. Of course, the Seminar corpus contains only four attributes, and for the two types where the theoretical risk of confusing them seems highest (STIME and ETIME), extraction performance is very high (99.3% and 97.1% F-measure, respectively, cf. Sec. 17.2). This reduces the potential for type confusions.

In this specific aspect, the Corporate Acquisitions corpus is very different, as we will see when we now turn to Figures 21.3 and 21.4. Here, pure *wrong type* mistakes (where, except for the type confusion, the prediction would have been correct) are with 13% the second most important kind of mistake combination (Fig. 21.3(a)). *Wrong type* mistakes also occur frequently in combination with other mistake types, resulting in a total of 18.2% of all mistake types (Fig. 21.4). Due to the high number of combinations this causes, we had to split Fig. 21.3 in two parts: Fig. 21.3(b) shows the 5.1% slide of less frequent combinations subsumed as *Other combinations* in Fig. 21.3(a). Again,



(a) Most Frequent Combinations



(b) Other Combinations

Figure 21.3: Acquisitions Corpus: Mistakes Combinations

mistake combinations occurring less than once per test run have been combined as *Others* (in Fig. 21.3(b)).

Except for the high frequency of type confusion mistakes, the distribution of mistakes in the Acquisitions corpus is similar to the Seminar corpus. Again, *completely missing* answer keys are the largest problem (causing 60% of all mistakes). *Completely spurious* predictions are far less frequent (11%), but still more important that the different kinds of boundary mistakes and *ignored* predictions. Misplaced boundaries are involved in 14% of all mistakes, when counting them all together; again, correctly identifying the end of attribute values is more of a problem than locating the start. *Ignored* predictions, however, are a far lesser problem in this corpus, and when they occur, it is often in combination with type confusions or boundary mistakes—only 1.8% of all mistakes are "pure" ignored predictions.

In the next section, we will look at the type confusion mistakes which cause so many problems in this corpus.

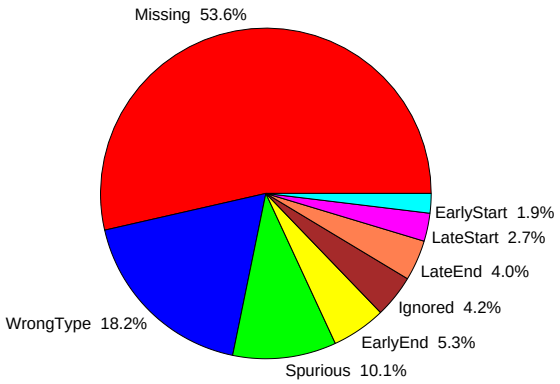


Figure 21.4: Acquisitions Corpus: Distribution of Mistake Types

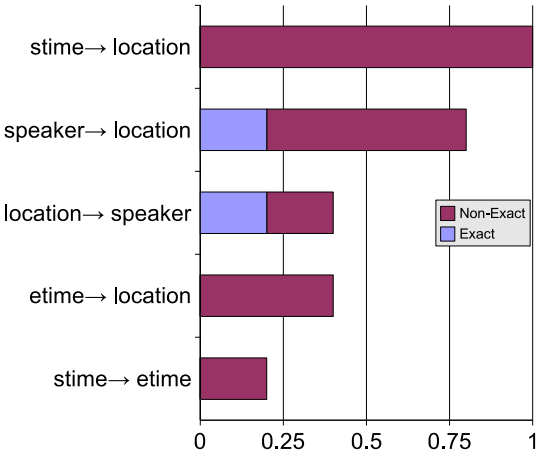


Figure 21.5: Seminar Corpus: Confusion Matrix (expected type→predicted type)

21.3 Type Confusion

Figures 21.5 and 21.6 show the confusion matrices for the type confusion errors that occur in the two corpora. The x-axis shows the average number of confusions per test run. Exact confusions (where the prediction would have been correct except for the wrong type) are shown in blue, while non-exact ones (where the confusion occur in combination with other mistakes, e.g. misplaced boundaries) are shown in purple.

As stated above, type confusion errors are almost irrelevant for the Seminar corpus (Fig. 21.5). In all five test runs, there are only two exact confusions: once, a SPEAKER is thought to be a LOCATION, and once, vice versa, a LOCATION is extracted as a SPEAKER.[1] Also, there are some non-exact confusions between other attributes, but none occur more than once on average. Interestingly, there is only a single non-exact confusion between an STIME (start time) and an ETIME (end time)—we would have supposed that the algorithm had more problems to differentiate between these two types, since both are times, but this is not the case.

[1] Since the chart shows the average over the five test runs, a single mistake shows up as 0.2.

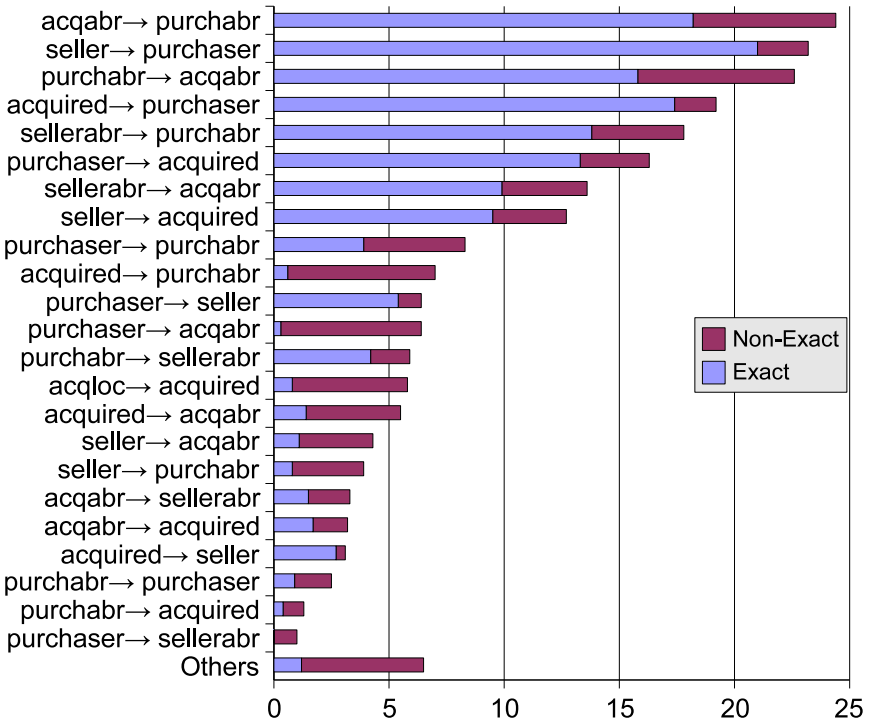


Figure 21.6: Acquisitions Corpus: Confusion Matrix (expected type→predicted type)

Figure 21.6 looks very different, due to the large number of *wrong type* mistakes in the Acquisitions corpus. Here we see a large number of confusion types, several of which occur more than 20 times on average per test run (i.e., more than 200 times in all ten test runs). Confusion types occurring less than once on average have been combined as *Other*.

More interesting than the mere number of confusions are the types of confusions that happen most frequently. In the last chapter (Sec. 20.2.2), we conjectured that the differentiation between long and short (abbreviated) names of the three kinds of companies to extract (ACQUIRED, PURCHASER, SELLER) would be especially tricky. However, the chart shows that this is not the case—instead, it is *logical* confusions about the role that a company plays in a transaction that cause most problems. The eight most frequent confusions types are all of this logical kind: the short name of the acquired company is thought to be the short name of the purchaser (range 1, most frequently) or vice versa (range 3), the long name of the SELLER (range 2) or of the ACQUIRED company (range 4) is thought to be the long name of the PURCHASER, etc.

The first confusion between a long and a short form (PURCHASER extracted as PURCHABR, i.e. the abbreviated form) follows only on range 9, with an average frequency of about 8—one third of the average frequency of the most important logical confusion (24). Less surprisingly, almost all cases of confusions involve two of the six company attributes to recognize; in some cases, the location of the acquired company is thought to be the company itself (ACQLOC→ACQUIRED, range 14).

In the following section, we will manually inspect the mistakes that occurred, to

learn more about the causes of such logical confusions, and of the encountered mistakes in general.

21.4 Additional Manual Analysis

In many cases, looking at the text does not provide any additional insight into the cause of the mistake. In other cases, however, we can detect some interesting patterns.

21.4.1 Seminar Corpus

Completely Spurious Predictions

Completely spurious predictions tend to be of the correct "logical type" (i.e., the supertype proposed in Chap. 14). Many spurious SPEAKER predictions are the names of persons, many spurious STIME and ETIME predictions are indeed time expressions. Also (though less generally) LOCATIONs often are actual locations, just not of the talk in question.

If no SPEAKER is given in a seminar announcement, the algorithm tends to extract a person who is somehow involved with the talk. For examples, *"Karen Brean"* (who is probably the organizer of the talk) is extracted from the following phrase:

> I got a call from Karen Brean today asking me to let everyone who was in the urbanlab that the architectural students are doing their jury presentations on Friday, May 5.

In another announcement without a SPEAKER, *"J. Ruppert"* is extracted from "Joint work with J. Ruppert".

Another such case is the announcement of a project presentation from "Group Members: David Cooke, Molly FitzGerald" (and others). No SPEAKERs are annotated, but the algorithm proposes *"David Cooke"*, the first group member, as SPEAKER (which seems indeed a reasonable choice).

When a SPEAKER is given, the system sometimes confuses the person organizing or inviting to the talk with the actual SPEAKER; e.g., *"Joe W. Trotter"* is extracted from an announcement signed "Sincerely, / / Joe W. Trotter / Professor of History" (where "/" indicates a line-break). In another text, *"Daniel Stodolsky"*, who invites to a talk given by **"Alessandro Forin"**, is extracted as SPEAKER. In both cases, the actual SPEAKER is found too, but is *ignored* due to a slightly lower probability.

Similar "logical" confusions occur with other attributes as well, e.g. in the phrase "Time: <**stime**>3:45</**stime**> (Refreshments at 3:30)", *"3:30"* is proposed as STIME (start time), while the actual STIME is considered less probable and hence *ignored*.

In the following sentences, the system extracts the underlined words as spurious LOCATIONs (which really are locations, just not of the talk announced):

> Copies of the articles will be made available at the Reserve Desk of Hunt Library and at the Special Projects Office of the Associate Provost in Warner Hall 419.

> Full information is posted on the bulletin board behind the monitor's desk in the Career Services Reading Room.

In a few cases, mistakes occur due to incomplete tagging of the corpus. For example, the system proposes *"Professor Paul"* as a SPEAKER while **"Professor Manfred Paul"** would have been correct; and *"Professor Katz"* instead of **"Professor Randy Katz"**. In both cases, the extracted attribute values refers to the same person, hence strictly speaking, they should have been correct too, but they are considered mistakes since they are not annotated in the corpus. (Again, the full names are also found by the algorithm, but they are *ignored* as less probable.)

Completely Missing Answer Keys

Completely missing answer keys are often placed in an unusual environment which makes them harder to detect for the algorithm, such an the phrase:

> The talk by <**speaker**>Max Henrion</**speaker**> has been moved to <**location**>Porter Hall 7A</**location**>

Usually, the texts in the corpus announce a new talk, while in this case the announcement reports the relocation of a talk which had (apparently) been announced before. This unusual phrasing causes the algorithm to miss both attribute values (LOCATION and SPEAKER), probably due to the lack of similar training data.

In the following example, a very unusual verb ("provides viewpoints") is used for introducing the SPEAKER, who moreover is separated from the verb by a long relative clause:

> <**speaker**>Jeannine Amber</**speaker**>, an African American Jewish freelance writer in New York City provides viewpoints on [...]

Similarly, it is probably the long distance between noun and verb that prevents the SPEAKER from being extracted from the following sentence (all other answer keys are found correctly):

> On Wednesday, November 10 at <**stime**>4:30 p.m</**stime**>. <**speaker**>Paula Rayman</**speaker**>, Associate Professor of Sociology and Director, Pathways for Women in Science Project at Wellesley College will speak at <**location**>1175 Benedum Hall, University of Pittsburgh.</**location**>

In other cases, the answer keys themselves deviate strongly from the usual form of values of this attribute; for example, the unusually informal LOCATION **"Fil's office"**; or the SPEAKER's name rendered as **"R A V I K I R A N"** (nine separate tokens).

Frequently missed are very long attribute values, for example LOCATIONs such as:

- **"Mellon Institute, 3rd Floor Conference Room"**
- **"La Roche College, 9000 Babcock Blvd."**
- **"Hamburg Hall, H. John Heinz III School of Public Policy and Management"**

– **"room 261 of GSIA in the new building"**

We will later see that both recall and precision reached by the algorithm generally fall when the length of attribute values increases (Sec. 21.5).

Early End

A frequent source for early ending predictions are unusual tokens within an attribute value, such as commas or parentheses (since these tokens usually occur after the end of attribute values). For example, SPEAKERs proposed by the system include *"Eric H. Nyberg"* instead of **"Eric H. Nyberg, 3rd"**, and *"Joel S. Birnbaum"* instead of **"Joel S. Birnbaum, Ph.D"**.

Similar problems occur with LOCATIONs, e.g. *"DH 3313"* instead of **"DH 3313 (large conference room"** (sic—trailing punctuation characters are usually not annotated as part of attribute values, but see below), or *"Main Auditorium (1st Floor"* instead of **"Main Auditorium (1st Floor) GSIA"**.

Another cause of *early end* mistakes are tagging inconsistencies regarding trailing punctuation in attribute values. In the Seminar corpus, times ending in "p.m." or "a.m." are usually (and surprisingly) annotated *without* the trailing dot (**"3:30 p.m"** instead of "3:30 p.m."), and LOCATIONs with parenthetical explanations are annotated without the closing parenthesis (see example in the preceding paragraph). But there are exceptions where the trailing punctuation *is* included. This leads to several erroneous predictions such as *"4:30 p.m"* or *"ITC Lecture Room (Rm 279"* where the expected answer is **"4:30 p.m."** / **"ITC Lecture Room (Rm 279)"**—unsurprisingly, while the system is able to learn the general rule, it fails to learn the exception.

Another mistake is caused by a formatting error: instead of **"Carnegie Conference Room, Warner Hal / l"** (with an accidental line-break in the word *"Hall"*, rendered as "/"), the system extracts *"Carnegie Conference Room, Warner Hal"*, considering the end of line to mark the end of the attribute value.

Late End

The tagging inconsistencies mentioned in the previous section occasionally also cause the inverted kind of mistake, leading to predictions such as *"CMT conference room (BoM 109)"* instead of **"CMT conference room (BoM 109"**.

Other *late end* mistakes involve trailing punctuation as well, e.g., *"WeH 8220."* instead of **"WeH 8220"** and *"Doherty Hall ??"* instead of **"Doherty Hall"** (in the latter case, the question marks probably indicate that the location is not quite fixed, so extracting them as part of the LOCATION attribute value is indeed not unreasonable).

In a somewhat similar case, the system extracts *"Jonathan Caulkins (*)"* instead of **"Jonathan Caulkins"**, treating a footnote marker as part of the SPEAKER's name.

Occasionally, we find other trailing tokens (*"3:30 and"* instead of **"3:30"**, *"CMT red conference room 10-11am"* instead of **"CMT red conference room"**. In the latter case, the system erroneously combines LOCATION with STIME and ETIME (however, the latter attributes are extracted correctly from another reference in the text).

Late Start

Problems recognizing the start of attribute values are rarer than those recognizing the end. Sometimes the title introducing a SPEAKER's name is missed or extracted only partially (*"Professor Harold L. Alexander"* instead of **"Asst. Professor Harold L. Alexander"**). In another case, the name itself is clipped (*"Claude Latombe"* instead of **"Jean-Claude Latombe"**).

In case of LOCATIONs, the algorithm sometimes overlooks room numbers at the start of attribute values (*"Wean Hall"* instead of **"623 Wean Hall"**, *"Mellon Institute"* instead of **"448 Mellon Institute"**); or it misses the first part of a long (two-line) LOCATION extraction (*"EPP Conference Room"* instead of **"129 Baker hall / EPP Conference Room"**).

Other Mistake Types

Early start mistakes occur very seldom in this corpus, and usually in combination with other errors.

Ignored prediction mistakes can only occur if there is another prediction that is considered more likely but turns out to be wrong. They have already been discussed above, in the section on spurious predictions.

Wrong type mistakes occur almost never in the Seminar corpus. For analyzing them, we will turn the Acquisitions corpus, where this type of mistake is especially frequent.

21.4.2 Acquisitions Corpus

In the Acquisitions corpus, we have only inspected the type confusion mistakes, since this is such a frequent kind of mistake in this corpus and we have already inspected the other mistake types in the context of the Seminar corpus.

Earlier (in Sec. 21.3), we had already noted—with some surprise—that the most frequent type confusions are logical (e.g., between SELLER and PURCHASER), not between short and long versions of the same name. By looking at the context of the mistakes that occur, we can now identify some reasons for this behavior.

Unspecific or Vague Context

One typical cause of confusions are sentences that contain general company-related information, such as:

> Headquartered in Somerset, N.J., <**acqabr**>PMS</**acqabr**>[2] reported over 70 mln dlrs in revenues in its last fiscal year [...]

Such statements, which are typical in press releases, occur quite frequently in the newspaper articles comprising the corpus. They provide no clues about the role of the company in the reported acquisition, making extraction almost a guessing game.

Various phrases report on the status of merger talks without providing logical clues about the roles of the companies involved:

[2] Extracted as PURCHABR.

<acqabr>NORCROS</acqabr>[3] BREAKS OFF MERGER TALKS WITH <purchabr>WILLIAMS</purchabr>[4]

<acqabr>ROSPATCH</acqabr>[5] <RPCH>; TO RESPOND TO <purchabr>DIAGNOSTIC</purchabr>[6]

The roles of purchaser and acquired company could just as well be switched in such cases, hence it is not surprising that the algorithm often fails to assign them correctly (or to extract them at all, if none of the rival attributes is considered sufficiently likely).

Lack of specific context is a problem in other cases as well:

<acquired>Norcros <NCRO.L> Plc</acquired>[7] said it has <status>no intention of proceeding any further</status> with talks on <purchaser>Williams Holdings Plc</purchaser>'s[8] suggestion that there would be benefits arising from a merger between the two groups.

<acqabr>Southwest</acqabr>[9] currently has 12.3 mln shares outstanding.

<acqabr>Advanced</acqabr>[10] said <purchabr>Sterling</purchabr>'s[11] board has decided not to enter the nicotine product market.

While these sentences contain subtle hints about what is going on, there are probably not enough similar examples in the training data to learn their meaning.

<acquired>Trans World Airlines Inc</acquired>[12] said chairman <purchaser>Carl C. Icahn</purchaser>[13] has <status>withdrawn his proposal</status> to acquire the <acqabr>TWA</acqabr>[14] shares [...]

In this case, only the fact that **"TWA"** is an acronym of **"Trans World Airlines"** makes it clear that the (not-)ACQUIRED companies issues this statement and not the purchasing company, but acronym matching is not part of our system.

Logical Confusions

In other cases, logical confusions occur, though for a human observer the answer is clear:

[3] Extracted as PURCHABR.
[4] Completely missing.
[5] Extracted as PURCHABR.
[6] Completely missing.
[7] Extracted as PURCHASER.
[8] System extracts only *"Williams"* as candidate PURCHABR.
[9] Extracted as PURCHABR.
[10] Extracted as PURCHABR.
[11] Completely missing.
[12] Extracted as PURCHASER.
[13] Recognized correctly but *ignored* since the other PURCHASER prediction is considered more likely.
[14] Extracted correctly.

> <**seller**>Butler Manufacturing Co</**seller**>[15] said it completed sale of its <**acquired**>Livestock Systems</**acquired**> division[16] [...]

> <**seller**>Synalloy Corp</**seller**>[17] said it has <**status**>ended talks</**status**> on the sale of its <**acquired**>Blackman Uhler Chemical Division</**acquired**>[18] shares [...]

Confusions such as these are probably caused by the fact that it more frequently the PURCHASER who is issuing such statements, and the clues in the context are too irregular or to far away for the algorithm to learn the distinction.

> <**acqabr**>AMERICAN DYNAMICS</**acqabr**>[19] <AMDC> TO SELL 51 PCT STAKE

A possible reason for this mistake is that "to buy" and related verbs are more common in such a context and the algorithm lacks sufficient training data to learn how to differ between such verbs and "to sell" (there are 60 instances of "to acquire", "to buy", and "to purchase" after the names of companies, but only 23 instances of "to sell" and no obvious synonyms). Also, verb phrases such as "to sell" can refer both to the company being acquired or to be selling company, hence they fail to provide an unambiguous context.

> <**acquired**>Foote Mineral Co</**acquired**>[20] said it has signed a <**status**>letter of intent</**status**> to merge into <<**purchaser**>Rio Tinto-Zinc Corp PLC</**purchaser**>[21]> for cash.

While the corpus contains three examples of "to merge into PURCHASER|PURCHABR", the similar (and more frequent) formulation "to merge with" is ambiguous—examples such as "<**purchaser**>Hughes Tool Co</**purchaser**> Chairman W.A. Kistler said its counter proposal to merge with <**acquired**>Baker International Corp</**acquired**> was still <**status**>under consideration</**status**>" might have been the cause of this mistake.

21.5 Length Analysis

An assumption we have voiced before is that longer attribute values are more difficult for the extraction algorithm. To probe this assumption, we have calculated the usual metrics (precision, recall, F-measure) separately based on the number of tokens each attribute value contains.[22]

[15] Extracted as PURCHASER.
[16] System extracts *"Livestock Systems division"* as ACQUIRED (*late end*).
[17] Extracted as PURCHASER.
[18] Extracted correctly.
[19] Extracted as PURCHABR.
[20] Extracted as PURCHASER.
[21] Extracted as ACQUIRED.
[22] Tokenization has been treated in Sec. 12.3.

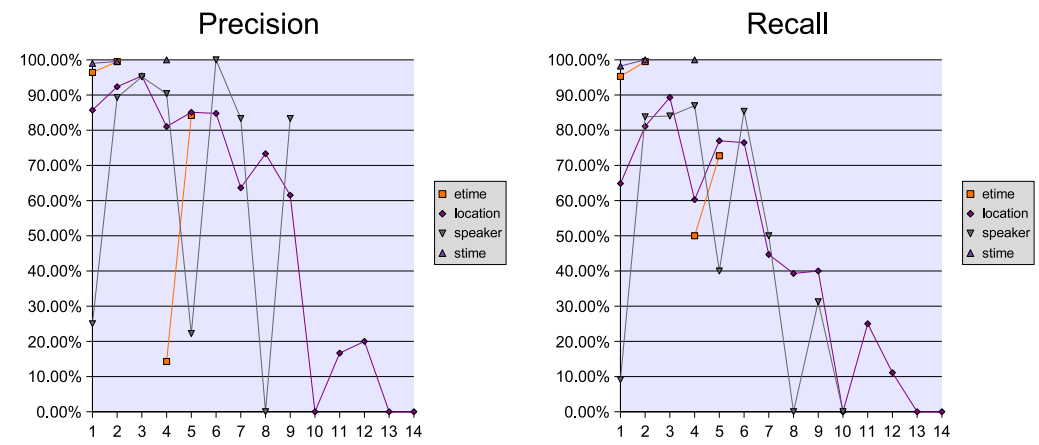


Figure 21.7: Seminar Corpus: Precision and Recall by Token Length

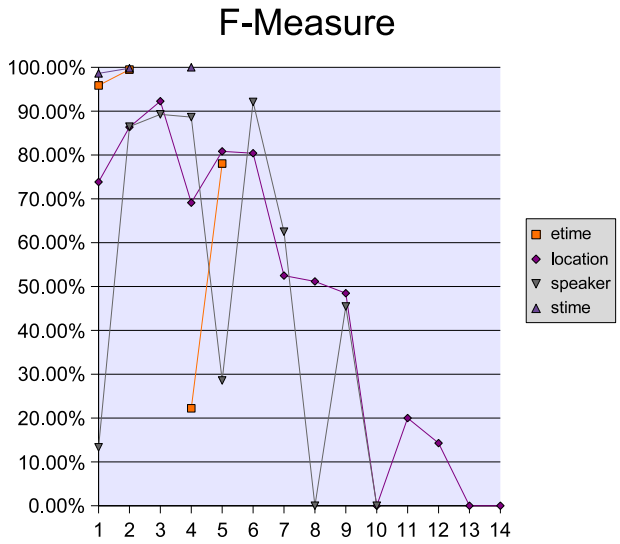


Figure 21.8: Seminar Corpus: F-Measure by Token Length

Figure 21.7 shows how precision and recall of the Seminar corpus attributes develop for values of increasing lengths; Fig. 21.8 shows their harmonic mean, the F-measure. From these figures we can see a general tendency of both precision and recall to fall as attribute values get longer, but it is hard to get a clear picture. Especially for the precision, there are various outliers where values jump erratically, typically when there are very few answer keys for an attribute/length combination. This is the case, for example, for ETIMEs (end times) of length 4 and SPEAKERs of length 5 and 8—see Table 21.1 for the distribution of answer keys of different lengths.

To allow us a clearer picture, Fig. 21.9 shows the *weighted average* (cf. Sec. 15.3) of the precision, recall, and F-measure values: for attribute values of each length, the different attributes are weighted by the relative number of answer keys of this length.

Tokens	etime	location	speaker	stime
1	29.5%	3.2%	1.1%	41.9%
2	66.4%	28.1%	63.8%	56.6%
3	–	32.1%	9.3%	–
4	0.3%	6.7%	18.1%	1.5%
5	3.8%	10.8%	0.5%	–
6	–	8.7%	4.0%	–
7	–	4.0%	1.0%	–
8	–	2.4%	0.5%	–
9	–	1.7%	1.6%	–
10	–	0.9%	0.1%	–
11	–	0.3%	–	–
12	–	0.8%	–	–
13	–	0.0%	–	–
14	–	0.2%	–	–

Table 21.1: Seminar Corpus: Length Distribution of Answer Keys

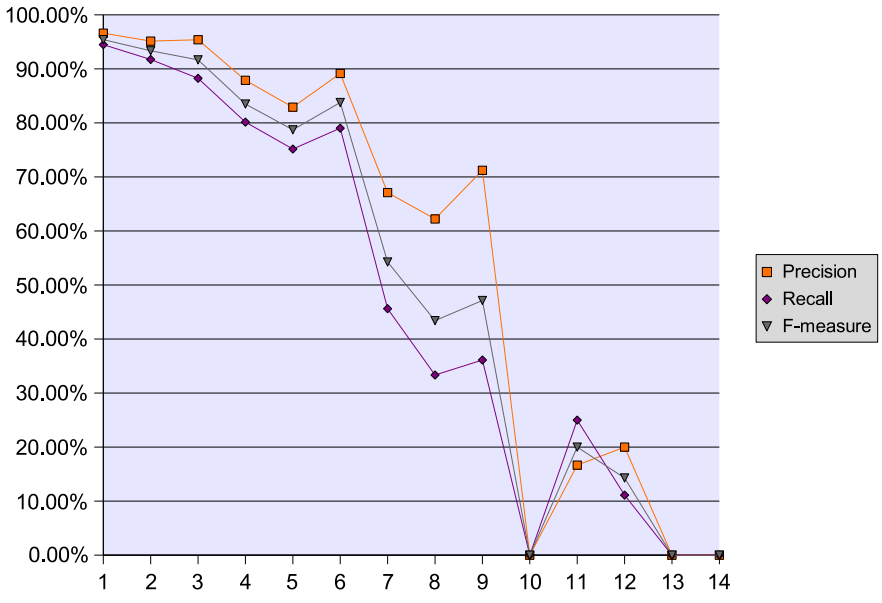


Figure 21.9: Seminar Corpus: Weighted Averages by Token Length

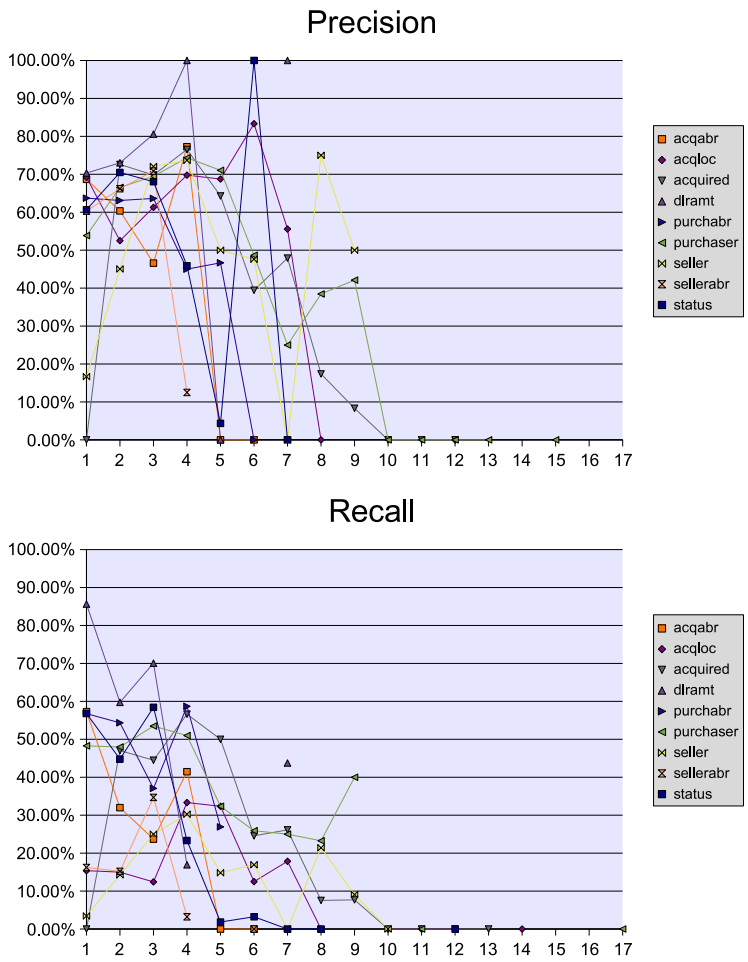


Figure 21.10: Acquisitions Corpus: Precision and Recall by Token Length

This shows clearly, that, on average, both precision and recall fall almost continuously for increasing token counts. Values of neighboring token lengths are often similar, but the general tendency is obvious. The average F-measure is 92–95% for attribute values with 1–3 tokens, 79–84% for 4–6 tokens, and 44–54% for 7-9 tokens. The algorithm fails to correctly extract *any* attribute values with 10 or with 13 or more tokens; after the sudden drop at 10, it recovers again and it able to correctly recognize a few of the LOCATIONs (the only attribute where some values are this long) with 11 or 12 tokens (about 15–20% F-measure).

Figures 21.10 and 21.11 show the individual metrics for the Acquisitions corpus; their weighted average is shown in Fig. 21.12 (Table 21.2 shows the distribution of answer keys of different lengths).

Here, we do not see a clear tendency for short attribute values with 1–4 tokens—in fact, the precision increases slightly up to this value. For attribute values with 5 or more tokens, both precision and recall fall dramatically. The algorithm fails to extract *any* values with 10 or more tokens, though there are answer keys with up to 17 tokens in the corpus. (Note that there are neither answer keys nor predictions with 16 tokens,

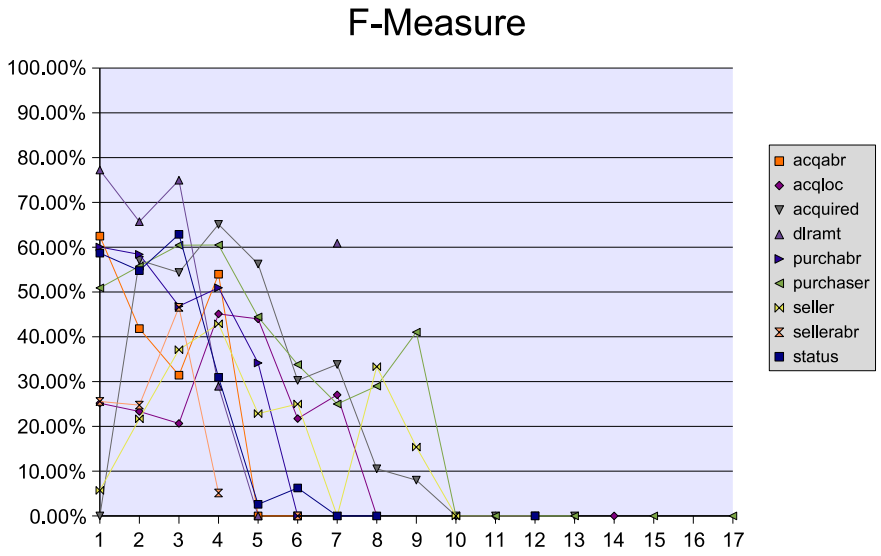


Figure 21.11: Acquisitions Corpus: F-Measure by Token Length

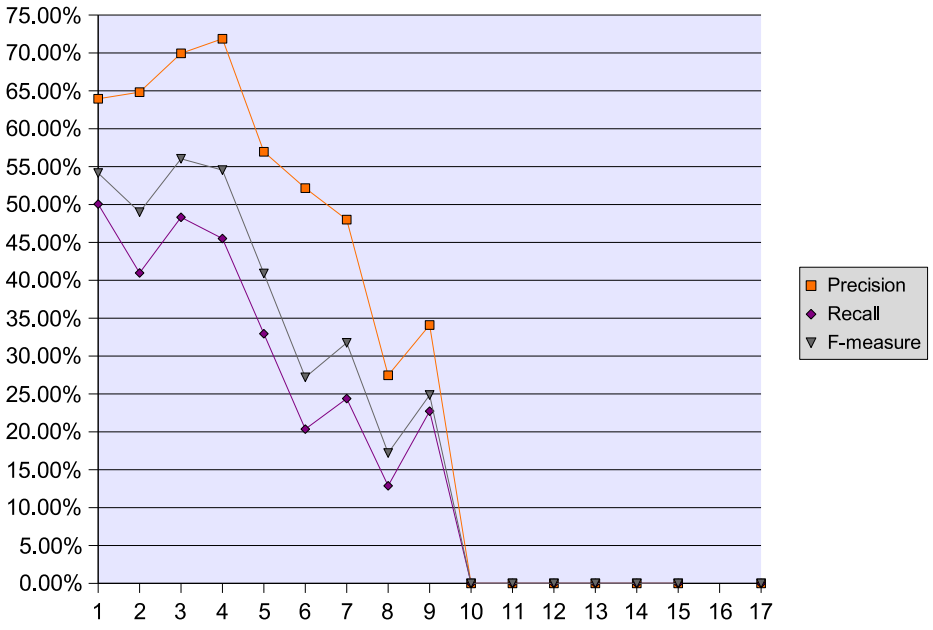


Figure 21.12: Acquisitions Corpus: Weighted Averages by Token Length

Tokens	acqabr	acqloc	acquired	dlramt	purchabr	purchaser	seller	sellerabr	status
1	63.9%	39.1%	1.4%	16.5%	62.1%	1.1%	2.5%	57.0%	41.8%
2	25.1%	15.6%	21.6%	23.5%	26.8%	28.1%	24.6%	33.6%	12.7%
3	7.9%	17.1%	39.8%	53.8%	7.6%	37.7%	44.2%	5.3%	35.9%
4	1.9%	10.0%	19.8%	5.0%	2.1%	18.5%	15.9%	3.4%	5.3%
5	1.0%	7.6%	7.3%	–	1.2%	6.1%	4.6%	–	2.4%
6	0.2%	4.5%	3.8%	–	–	5.1%	5.1%	0.7%	1.4%
7	–	3.1%	3.0%	1.2%	–	0.7%	0.7%	–	0.2%
8	–	1.3%	1.8%	–	0.2%	1.6%	1.2%	–	0.2%
9	–	–	0.4%	–	–	0.7%	0.9%	–	–
10	–	0.6%	0.5%	–	–	–	0.3%	–	–
11	–	0.3%	0.2%	–	–	0.1%	–	–	–
12	–	0.3%	0.2%	–	–	–	–	–	0.2%
13	–	–	0.1%	–	–	–	–	–	–
14	–	0.4%	–	–	–	–	–	–	–
15	–	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	–	–
17	–	–	–	–	–	0.3%	–	–	–

Table 21.2: Acquisitions Corpus: Length Distribution of Answer Keys

hence precision and recall are undefined and not 0 for this number.)

For both corpora, we also see that the general tendency of our algorithm to favor precision over recall holds for attribute values of any length.

The results we can draw from this analysis will be discussed in the next and final chapter.