Context Selection For Linguistic Data Fusion

Nicholas Morizio
Lockheed Martin Advanced Technology Laboratories
Cherry Hill, NJ, U.S.A.
Email: nmorizio@atl.lmco.com

Abstract—Reports generated by soldiers are common in time-critical military environments. Data fusion systems that attempt to process those reports must maintain the context for each set of observations to avoid inaccurate state estimates. This paper analyzes the selection and assignment of topical context under a Bayesian methodology. We present several techniques to decrease the hypothesis space and heuristics that apply specifically to military reporting environments. Using a data set consisting of semi-structured reports, we show that this approach allows accurate assignment of topical context even when the context is only implied rather than given explicitly.

Keywords: Context, soft data fusion, correlation, reporting.

I. INTRODUCTION

Human-generated reporting is common in a variety of military settings. Such reports are often semi-structured; meta-information is encoded in a structured way and other important content is encoded as free text. Aggregating these reports over long periods of time to provide a consolidated situation estimate is a largely manual process in current operations. Large quantities of received reports make this process tedious, time consuming, and error-prone. Therefore, automated data fusion systems to process these reports are desired. This type of data fusion is sometimes characterized as soft data fusion, and the resulting state estimate can take a graphical form such as a social network.

This paper presents an approach to the selection of context for observations in the reports. In general we define the context of a set of observations as the information that limits possible interpretations of the elements in the set. The context usually includes some spatial and temporal boundaries that are specified imprecisely, such as “this morning in southwestern Philadelphia”, as well as some topic for which the observations are salient. The context limits potential correlations with previously instantiated objects and so provides a gate mechanism for the data fusion process.

The context itself may be implicit or may be partially specified in meta-data associated with the report. Several contexts can be used within the same report. Also, contexts can be specified at varying levels of abstraction or generality. Because of this uncertainty over the actual context for a given observation, an approach dealing directly with the uncertainty is appropriate. Our focus in this paper is the uncertainty over the topical context for a given set of observations.

This is not to be confused with data fusion for context-aware computing, where the goal is to model and maintain the context of the user, which allows the fusion system to better allocate sensing and data fusion activities in an attempt to promote relevancy and accuracy of the results [1]. Our focus here is on the context of the reporter when choosing a particular expression for observations in a report. We call this the report context.

Report context also includes background information about the environment in which the reporting occurs. Certain assumptions will be made in reports based on implicit information not cited directly. This type of context can be specified in a domain ontology, with rules and facts implemented on a per application basis. We assume this approach for the remainder of this paper.

As a motivating example for the importance of topical context, consider the following extraction from a synthetic report:

ORANGE 14 OBSERVED GREEN 11 SHOT DOWN BY PROBABLE SA-10.

Without representing the context, ORANGE 14 could be any type that is capable of taking on the relation OBSERVED according to an ontology of entities and relationships of interest. This may be a large number of possible types, and given no prior preference towards certain types, the classification is largely ambiguous. This ambiguity will complicate correlation in the data fusion system since correlation metrics between ambiguous types must default to string comparison. If, via the same ontology, we are told that ORANGE 14 is equivalently represented as OR14 in this domain when the topic is air operations, we will be able to correctly correlate the two across separate reports even in the presence of varying syntax.

One option is to treat this as a text classification problem where the goal is to classify a set of observations as belonging to one topic in a set of topics. This problem has been studied extensively in the literature [2]. However, in practice, sets of observations in the same report can have disjoint contexts, and each individual observation may have a slightly different context from its neighbors. Instead, our approach determines where the context changes and usually finds the most likely derivative topics for each observation.

Let \( \omega \in A \) denote a topic of any context that is defined in an ontology for a given domain. Also, let \( \omega \in \Omega \) denote an assignment of context topics to observations. The posterior probability of \( \omega \) is given by Bayes rule:

\[
P(\omega | O) = \alpha P(O | \omega) P(\omega)
\]

(1)

Let \( T(o_n) \) denote the topic assigned to observation \( o_n \). The

...
probability that $T(o_n) = a$ is given by:

$$P(T(o_n) = a | O) = \sum_i P(T(o_n) = a | O, \omega_i) = \sum_i P(T(o_n) = a | \omega_i) P(\omega_i | O)$$  \hspace{1cm} (2)$$

This paper addresses several difficulties in solving (2) efficiently. Section II-A deals with grouping observations under a common context, which assumes that context changes less frequently than observations appear in a report. This is often the case, especially when reports are brief and specific. Making use of this assumption greatly improves the tractability of the summation in (2). Section II-B continues this theme, describing how we use an ordering of the topics to further improve the tractability.

Section II-C deals with attempts to estimate $P(\omega)$. Assuming a uniform prior is a common approach; however, we achieve better results by taking advantage of the hierarchical nature of our model of context.

Finally in Section III we apply our approach to a synthetic data set consisting of reports that are representative of a type common in military environments. We compare results from several versions of our assignment algorithm. Section III-A discusses some limitations of our approach and considers an extension to make use of spatial and temporal constraints in the context when available.

II. ASSIGNING TOPICAL CONTEXT

Because (2) requires summing over an exponential number of hypotheses, an obvious goal is to reduce the set of hypotheses under consideration. We begin by examining the way context changes within the report types of interest.

A. Context grouping

We first note that when the context is implicit, it does not usually change in the report body\(^1\). This is because a proper cue is provided when the context changes. Such cues can be subtle, however, in that they often provide only a partially explicit new context. In our approach, a missed context change cue can be a large source of error. This is discussed in more detail in Section III-A. For now we assume that if no explicit context is given for a set of observations, then the set shares one common context. This effectively reduces the size of the hypothesis space, often by several orders of magnitude.

Next we note that within a context group only certain entities of interest are reported. Certain concepts, such as a target, are shared in many contexts, while others, such as a squadron, are limited to a narrower set of topics. In our ontology, relations have limited domains and ranges based on concept types, and those concept types are restricted to a set of contexts. These restrictions are chosen to specify the relations, concepts, and events of interest in the state estimate output from the data fusion system. Such restrictions simplify the computation of the likelihood of the observations given the hypothesis, the $P(O | \omega)$ term in (1).

Given the aggregated observations in the context grouping, we can make use of these restrictions along with the hierarchical nature of context topics to estimate the above likelihood.

B. Topic hierarchy

Let $C(a)$ denote the subtropics of $a$, which are specializations of the given topic, and let $R(a)$ denote the topical generalization of $a$. If $R(a) = \emptyset$ for any topic $a$, we set $R(a)$ to a special placeholder root, which is not a valid topic. The ordering imposed by $C(a)$ and $R(a)$ form a directed graph of all topics in the domain. When we add the restriction that no topic $a$ can have a specialization that is more general than $a$ according to a topological sort, the graph becomes acyclic. This restriction is less severe than it seems because if necessary we can add topics that are a combination of two or more topics as special nodes in the graph.

We say an assignment of topics to any two observations $o, o' \in O$ is inconsistent if and only if $o$ and $o'$ share an entity in common and no path exists from root that contains both $T(o)$ and $T(o')$. Further, any assignment to all observations $\omega$ is inconsistent if and only if there exist two or more observations in $\omega$ whose assignments are inconsistent.

Any hypothesis $\omega$ that contains an inconsistent assignment is implicitly assigned $P(O | \omega) = 0$. For the limitations imposed by this assumption, see Section III-A.

Next we denote an assignment of topic $a$ to $o$ as valid if and only if the relation $r$ in $o$ is defined under $a$ in our ontology. If this is the case, the concept types for entities in $o$ are restricted to those in the domain and range of $r$ under $a$. An assignment $\omega$ is invalid if and only if the assignment to each observation $o$ under $\omega$ is valid and each entity appearing in (potentially more than one) $o$ has a non-empty concept type set after all restrictions are applied. This guarantees that each valid hypothesis supports at least one interpretation for every entity in every observation.

\(^1\)Unfortunately, some assumptions on proper grammar and structure must be made for current natural language processing technology. Even human analysts have difficulty with poorly authored reports!
This rule prevents absurd\(^2\) assignments. For example, consider the ontology and report extracts in Figure 1. If the relation detected applies to GREEN 10 in one observation, GREEN 10 may be either a facility or an aircraft. However, if GREEN 10 flew in another observation, it is further restricted to an aircraft, because facility is not in the domain of flew in any context.

Any hypothesis \(\omega\) that is both consistent and valid given \(O\) is assigned a positive likelihood. In assigning likelihoods, we apply the principle of Ockham’s razor as a heuristic. That is, a more specific topic is considered more likely than a more general hypothesis. To enforce this heuristic, we weight each valid and consistent \(\omega\) according to the relative depth of its assigned topics in the graph imposed by \(C(a)\) and \(R(a)\). For a branch of depth \(d\) and a topic in that branch of depth \(D(a)\), we weight the unnormalized likelihood by \(\gamma^{d-D(a)}\) where \(0 < \gamma < 1\) is a system parameter that specifies our preference for more specific topics. This weight is applied to each context group, giving a likelihood for each \(\omega \in \Omega\).

We next turn our attention to the prior probability of a topical context assignment \(\omega\).

C. Prior probability of a context

The prior probability of a given hypothesis concerning a topical context assignment will vary by domain. Because such priors cannot be exactly determined, several approaches appear in the literature [3]. The most simple approach is the uniform prior, where

\[
P(\omega) = \frac{1}{|\Omega|}
\]

This approach provides reasonable results in practice when the set of possible contexts is not too complex. However, when the number of potential topics becomes large, a non-uniform prior may greatly improve the accuracy of the system.

For domains where the size of the reports is small and context is often implicit, learning the priors from a statistically significant sample set is a good option. However, access to such sample sets is often limited, and the model of priors needs to be re-trained whenever the ontology changes or, potentially, for each application of the fusion system in a new operating environment, which is impractical.

Instead, we provide a heuristic to model the prior likelihood of a hypothesis, which has shown better results than a strictly uniform prior in our experiments. We begin by observing that it is uncommon for context to vary wildly in a given report. That is, topics for observations in separate context groups tend to be closely related. In terms of our topic graph, this means they are more likely to have common ancestors.

We denote the number of context groups in \(\omega\) as \(|\omega|\). Let \(H(a) \leftarrow \{\text{root,} \ldots, a\}\) denote the set of nodes on the path from root to \(a\) in the graph imposed by \(C(a)\) and \(R(a)\). The common ancestor score \(S(\omega)\) for a hypothesis \(\omega\) is given by

\[
S(\omega) = |\omega|^{-1} \sum_{a \in \omega, b \in \omega, a \neq b} \frac{|H(a) \cap H(b)|}{|H(a) \cup H(b)|}
\]

Using the above heuristic and assuming a uniform prior for simple hypothesis with only one context, the prior probability of a hypothesis \(\omega\) is given by

\[
P(\omega) = \alpha \left( \frac{|A|}{|\Omega|} \right)^{-1} \quad \text{if } |\omega| = 1,
\]

\[
P(\omega) = \alpha \left( \frac{1}{|\Omega|} + \frac{1}{|\omega|} S(\omega) \right)^{-1} \quad \text{if } |\omega| > 1.
\]

where \(\alpha\) is a normalizing constant ensuring the prior sums to one.

In Section III we compare results using the common ancestor score (CAS) to results using a uniform prior.

III. EXPERIMENTAL RESULTS AND ANALYSIS

We have applied an implementation of our approach to a synthetic data set consisting of small reports with one to ten observations per report. The ontology modeled roughly 30 context topics with an average graph depth of 5. In addition, the ontology contained 104 concept types including physical (such as person) and abstract (such as agreement) concepts, with 23 relations including several non-binary relation types. The content covered several topics that might be found in typical military reports. We compared the topics selected by our implementation against the true topics as labeled by a user.

Table I shows the results of our implementation when run on this data set. The CAS prior provides better performance than the uniform prior in all cases. For our sample set, a \(\gamma\) of 0.5 provided the best performance among the test cases. Although the best value for \(\gamma\) would vary with the ontology, in general we expect a moderate value from 0.3-0.7 to give the best results.

<table>
<thead>
<tr>
<th>Prior</th>
<th>(\gamma)</th>
<th>% Correct</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>0.1</td>
<td>61.3</td>
<td>69.8</td>
</tr>
<tr>
<td>uniform</td>
<td>0.5</td>
<td>75.6</td>
<td>71.3</td>
</tr>
<tr>
<td>uniform</td>
<td>0.8</td>
<td>68.4</td>
<td>70.9</td>
</tr>
<tr>
<td>CAS</td>
<td>0.1</td>
<td>78.2</td>
<td>76.8</td>
</tr>
<tr>
<td>CAS</td>
<td>0.5</td>
<td>84.8</td>
<td>77.0</td>
</tr>
<tr>
<td>CAS</td>
<td>0.8</td>
<td>82.0</td>
<td>77.3</td>
</tr>
</tbody>
</table>

Since the context selection algorithm is just one part of our data fusion system, it does not need extremely high accuracy to be useful. Even so, the best case result of nearly 85% correct topic assignment shows promise for the problem of topical context assignment.

Figure 2 shows the computation time of each version of the algorithm as a function of the number of observations. From the plot we can see that our implementation scales well with the size of the input. Some portion of the time for each algorithm is spent performing other functions in the fusion
system; however, this accounts for only a small amount of the listed computation time.

Our implementation requires a natural language processing platform for parsing and tagging. For the results given here, we used OpenNLP, which is based on a maximum entropy framework for natural language processing [4]. Although evaluating the accuracy of the NLP software was beyond the scope of our experiment, a casual examination was enough to note that the NLP tagging was often inaccurate. We would expect to obtain better results with increased NLP performance.

A. Limitations and extensions

Our approach aims for good results in practice and uses approximations and heuristics in place of a more formal and theoretically optimal formulation. It also depends on certain key assumptions which may not hold for reporting environments.

One important assumption is that context change can always be detected through various cues. However because of the subtleties of natural language and the difficulty of natural language processing, these can sometimes be missed. When a context change is missed, it forces all observations in the separate contexts to conform to the same topic and so the topic is forced to be very general. This leads to ambiguity in concept types, and so the data association stage of our fusion system may perform poorly on that report. If context change cannot be perfectly detected, the next best solution is a more robust data association phase that is forgiving of vague contexts.

Another assumption is that context changes relatively infrequently compared to the number of observations in a report. In the worst case, each observation has its own explicit context in a very large report. Thankfully, such reports are rarely observed in common military settings. In cases where such reports are common, our approach is likely to perform poorly.

Finally, we choose to assign a zero probability to inconsistent or invalid hypothesis because this greatly reduces the size of the hypothesis space. This works well when the underlying ontology is complete and accurate. When using an ontology that does not accurately model the contexts, concepts, and relations in the domain, the correct hypothesis may be inconsistent or invalid, leading to poor results. In general, this potential problem can be avoided by sacrificing detail in the ontology; however, the trade-off is less-detailed state estimate outputs from the data fusion system.

As noted in the introduction, spatial and temporal bounds are often important components of the context. While a location or time may be given explicitly as part of an observation, the time frame or area in which there is a potential to observe is often implicit given the mission or role of the reporter. Although our implementation extracts and maintains these context features, the inference process for context topics does not use them as evidence.

In many cases we would expect that the topic could be further conditioned on the area of operation and the time frame of the mission. In general given a set of features for representing the context, our assignment hypothesis \( \omega \) could include an assignment to each context feature for each observation. However since, as noted, these features cannot be assumed independent, that approach would greatly increase the complexity of computing the likelihood function.

It is important to note that the data association phase of the fusion system will attempt to correlate events based on the full context of the observations, so providing a complete and accurate context is critical. Whether the potential increase in accuracy obtained by computing a joint likelihood over all context features would be worth the additional computational cost and challenge remains unanswered.

IV. SUMMARY AND FUTURE WORK

We have described an approach to selecting and assigning context topics to observations in semi-structured reports. This is a key component of our human-generated report fusion system. We have shown that this approach works on data similar to a common type of military report and that with certain assumptions about the structure of those reports, the computational costs are not prohibitive.

In the future, we would like to extend this approach to other types of context features and relax some of the assumptions we made in this implementation. We also continue to work toward a full human-generated report fusion system that is robust and reliable. Solving the context assignment problem is only one step in a full data fusion capability. We hope to be able to evaluate this system in a full military environment after further refinement.

REFERENCES