Did you mean A or B? Supporting Clarification Dialog for Entity Disambiguation

Anni Coden, Daniel Gruhl, Neal Lewis, Pablo N. Mendes

IBM Research, USA

Abstract. When interacting with a system, users often request information about an entity by specifying a string that may have multiple possible interpretations. Humans are quite good at recognizing when an ambiguity exists and resolving the ambiguity given contextual cues. This disambiguation task is more complicated in automated systems. As systems have more and more different entities and entity types available to them, they may better detect potential ambiguities (e.g., 'orange' as a color or fruit). However, it becomes harder to resolve entities automatically and effectively. In this position paper we discuss challenges in interacting with users to ask clarifying questions for entity identification. We propose three approaches, illustrating their strengths and weaknesses.

1 Introduction

Let us assume we are executing a Natural Language Processing (NLP) task with the help of a system which uses a knowledge base containing multiple entities and entity types. For example, assume that a user asks for information about an entity by entering a keyword (such as 'orange'). The meaning of this string is ambiguous: the system may have multiple entities known as 'orange' (e.g., a fruit, a company, a color and possible others). In some cases, the meaning of 'orange' may be disambiguated by qualifying it with the entity type (e.g., fruit or company). Other times the entity type alone may not be sufficient to disambiguate a term like 'apple'. Consider the sentence "Apple is quite popular for these cases studies" which could be found for instance business management course. Here one desires to not only differentiate between apple (a fruit) and apple (a company), but also differentiate between Apple Inc. (the computer company) and Apple Corps (the multi-media company founded by the Beatles).

Knowledge sources such as DBpedia [5] provide hundreds of thousands of entity types in the form of ontology classes, Wikipedia categories, etc. An entity type is often comprised of thousands of entities – e.g. there are more than 60,000 entities of type company. Meanwhile, a given string may have multiple possible meanings belonging to different entity types – e.g. the string 'apple' may refer to tens of entity types (e.g. plant, company, band, place, etc.) and even to several different entities of the same type in DBpedia (e.g. Apple Inc. and Apple Corps). This richness of knowledge sources comes at the cost of increasing the likelihood of ambiguity both with respect to entity type and entities.

Previous work has demonstrated effective ways to select candidate entities from DBpedia for a given string [9]. Often, it is possible to automatically decide on the correct interpretation for an ambiguous string [10, 3]. In other cases (e.g. for lack of context) either an automatic decision cannot be made or it is more advantageous to interact with the user to clarify the intended meaning. To effectively communicate with users, a system needs the ability to summarize its knowledge about possible entities given a string and concisely present it to the user in the form of a clarifying question.

In this paper we present three different methods for automatically generating such questions. We will highlight the main challenges and discuss pros and cons of the approaches.

2 Related Work

Schlangen [14] studied causes and strategies for requesting clarification in spoken dialog. Their fine-grained model of problems that may arise during the processing of a question included lexical ambiguity, but did not go into details of how to clarify ambiguity. Stoyanchev et al. [15] built a rule based system to generate clarification questions. In their experiments, the automatically generated questions performed better than a set of human generated questions. Rule-based systems have the disadvantage that they have to be manually adapted for new domains and new language patterns. Loos and Porzel [6] explore the use of hand-labeled data and ontological distance as methods of performing word sense disambiguation in speech recognition, and find that they both do well, although require a fair bit of human effort to develop relevant ontologies or score domain relevant training data. De Boni and Manandhar [4] developed an algorithm for clarification dialog recognition, in other words, to determine when a set of given questions are related and hence provide context for each other. They establish that clarification dialogues simplify the task of answer retrieval. In our case, we are trying to come up with the clarification questions for the specific case of ambiguity in one or more entity types in the original question.

There is a rich body of work on word sense disambiguation. Given the possible senses of a word, such systems assign the appropriate sense within the context. In general, such disambiguation methods are machine learning based with features chosen for the task. In the medical domain, the Unified Medical Language System [8], provides a rich ontology of terms and phrases with assigning different possible meanings to a word. For example the word 'ms' can refer to at least twelve quite different meanings. Previous work showed different types of documents (e.g. clinical text and biomedical literature) require different features to be used in machine learning algorithms to achieve the best disambiguation accuracy [13]. Such disambiguation systems are therefore costly to develop. Furthermore, even the best machine learning systems for a particular domain may find certain entities difficult to disambiguate, and therefore generate a low-confidence decision. For use cases that require high levels of precision (e.g. medical coding) our method can be applied to request human clarification when disambiguation systems are not very confident.

3 Entity Presentation in Natural Language Clarification Dialog

We investigated three classes of clarifying query (CQ) approaches: type-based, example-based and usage-based.

Type-based CQ – In some respects this is the most straightforward method. Suppose there are two entity types that have ‘orange’ as a member, and suppose furthermore that the labels of the types have semantically distinguishable names (e.g., a fruit type and a color type), then the system can use these labels to construct a clarifying query: “Do you mean ‘orange the fruit’ or ‘orange the color’?” Naturally, this method hinges on the availability of understandable and descriptive labels for the types. This method would be less helpful if the type labels are “noun” or “thing” or “type1234”. Furthermore such questions are not helpful to disambiguate entities within the same entity type.

Example-based CQ – In cases where the entity class labels are not available (or useful) the next easiest way to clarify the meaning is using other example entities that are similar to the possible meanings. Suppose the system is trying to clarify whether an entity of type fruit or color is meant. It might propose clarifying examples using other exemplars of each possible meaning: Do you mean ‘orange’ like banana and apple, or ‘orange’ like yellow? To do so the system needs to choose how many and which examples to provide. Our experience is that just showing the alphabetically first few members (or a random subset) of a group (e.g. an entity type) does not work well, especially in cases where a list of type members may run to the thousands.

Usage-based CQ – The last method we consider is to provide examples of natural language usage of a term in context. For example “Do you mean ‘orange’ as in ‘I’d like an orange juice.’ or ‘I love the 4G speed Orange now offers.’ ?” We can find the context by analyzing a large corpus where the entities of the target type occur many times. We look at these occurrences to identify contexts that are fairly specific to the type in question (e.g., fruits often occur in the pattern ‘`* juice`’.) This is difficult to do well but can be a very effective way to describe an entity: appropriate context snippets can capture nuances of meaning that even several examples cannot.

4 Qualitative Analysis

There are many methods one might use to “clarify” which entity is meant among entities of potential interest. We will focus on the three methods mentioned above and provide some qualitative analysis of challenges and opportunities with each approach.

4.1 Materials: Data Sets

For our preliminary analysis, we will restrict ourselves to entities of three entity types which have clearly very different meanings but contain overlapping terms: companies, colors and fruit. While our techniques apply also when the semantic meanings of the labels are quite similar and subjective (e.g., “warm colors” vs “sunset colors”), illustration of their pros and cons is simpler with pronounced differences in entity types.

The text for usage-based clarifying questions was (in general) extracted based on the ukWaC corpus [1]. Documents in this corpus are from a large web crawl focused on UK sites, so there are some regional nuances (e.g., some companies such as Orange S.A. are more commonly discussed in the UK than the US).

The lists of entities used for exampled-based clarifying questions were created by starting with a small seed set of members and expanding them with the ukWaC corpus using Concept Expansion[1]. This resulted in 131 companies, 124 colors and 66 fruits. These entity sets are clearly not “complete” but they are large enough to illustrate the approaches we are discussing. In practice, no entity type lists can be complete as new entities are always being added to the knowledge base.

4.2 Results

For the analysis in this section we will concentrate on a user query for ‘orange’ as our running examples, as it can be a fruit, a company, a color, etc. Questions are generated so that the user needs to choose among two alternatives: “Did you mean A or B?” Answers by the users will be ‘A’ if the first option is correct and ‘B’ if the second option is correct. If none of the alternatives is correct, the user will answer ‘None’ and the system will try again with another question.

Type-based clarification The generated clarifying question would be in the form:

Do you mean ‘orange’ as in a fruit or a company?

In cases such as this where there are clear and semantically meaningful entity type labels available this method is straightforward. However, this is not always the case, especially if the type lists are federated from multiple sources. For example, the label of a list derived from an optics textbook might be “590-620nm” (the wavelength of orange light). A list obtained from a design source might have technical names such as “warm colors” and “cool colors”. If the user does not know the meaning of the type names, the clarifying question will not be helpful. Lastly, some lists are generated from taxonomic downward closures or automatic clustering, and thus many have no names at all. In these cases a naïve type-based clarification question scheme breaks down, and alternatives have to be considered.

[1] `http://ibm.biz/WatsonCE`

It is worth noting that when a single phrase is legitimately ambiguous within a type (e.g., 'apple' as a company being the name of both a technology company and a multi-media company) the system will be challenged to differentiate them. That being said, given more fine-grained entity types (e.g., technology companies and recording companies) it can work to separate these. Therefore one of the challenges for this method is to determine the best granularity level to trade-off ease of understanding and distinguishability of an entity type.

Example-based clarification For this approach we wish to generate a small number (say three or four) examples from each potential interpretation to present to the user for clarification. The question would be of the form:

> *Do you mean 'orange' like mauve, lilac and pink, or 'orange' like IBM, Compaq and Hewlett Packard?*

We would like to choose examples that make the question easy to understand and apt to distinguish between the competing interpretations. The challenge is how to choose appropriate examples.

One effective way to select examples is to compute the distance between all term pairs in a type using a vector space model approach such as word2vec [11] and then for each member compute the sum of the distance to all other terms. Those with the smallest total sum are in a sense the "median" of the terms of that entity type. We can think of this as finding stereotypical terms from each type, which reflects our intuition of what would make good example terms for a clarifying question. In the above example, the stereotypical terms are mauve, lilac and pink – three colors that have less ambiguity than others such as coffee or orange. The same holds true for IBM, Compaq and Hewlett Packard with respect to companies. See Table 1 for more examples. In most of the peripheral cases one can see why the entity string might be a member of multiple types (e.g., brother is both a company and a family member, coffee is both a beverage and a color, etc.)

A simpler (and perhaps more intuitive) approach would be to select just the closest (e.g., by cosine distance in word2vec vector space) member of the type with regard to the target term (e.g. 'orange'). The examples generated in this manner turn out to be weighted towards those that are ambiguous in the same way the target term is. That is, suppose 'blackberry' belongs to the company, color and fruit types, 'blackberry' would end up close to 'orange' due to their shared ambiguity. Instead, we want terms that highlight the differences between the two possible interpretations. We refer to those "median" terms as being most central to the "meaning" of a type (e.g., apricots, cherries, plums are central to fruit). Terms that have multiple common meanings besides the one under consideration end up with a larger distance and thus are more "peripheral" (e.g., Sun and Brother, while companies, have other common usages).

Usage-based clarification Examples of sentence snippets including the ambiguous term provide an elegant solution to clarifying which entity is desired.

Type	**Central**	**Peripheral**
Colors	mauve, lilac, pink, taupe	coal, sage, bordeaux, coffee
Fruit	apricots, cherries, plums, melon	mulberry, berry, mandarin, orange
Companies	IBM, Compaq, Hewlet Packard, Lucent	Sun, Brother, Hughes, Myspace

Table 1. Examples of central (i.e. good) and peripheral (i.e. bad) terms from entity types. Note the "bads" have been filtered slightly to remove misspellings, etc.

The challenge is to automatically generate snippets that are both self-contained (i.e. contain enough information to be understandable) and helpful to distinguish between two competing interpretations of a term. We have developed a system that generates sentence snippets by first finding "appropriate" sentences in a source corpus and then extracting the clarifying phrases.

We start by leveraging Concept Expansion's pattern generation [2, 12]. By doing so we identify hundreds of patterns that have a high probability to contain an entity type in them. For example, for the type color:

– '`black / *`'
– '`and * in colour.`'
– '`of * vellum`'
– '`red, * and blue`'
– '`shades of * and`'
– '`* and purple`'

Patterns are scored as to the specificity of the context to the given type, and in the most selective patterns the * (in the list above) is replaced with the word of interest ('orange' in this case). We then query a corpus such as ukWaC for occurrences of such patterns (e.g., starting with 'black / orange'). If we get a match, the sentence in which it appears is selected. Otherwise the process is repeated with the next best pattern and so forth until a match is found. This algorithm yields the following sentence:

Orange and purple crayon streaks rainbowed across his briefcase surface.

While in this case the sentence can be used as is as a clarifying statement, for others (e.g., run on sentences that can go to half a page) it is helpful to select just the clause of interest. In addition a sentence can contain multiple mentions of words of interest with different meanings for each of them: e.g "I like the orange shirt; it reminds me to eat an orange." To determine an appropriate snippet we apply the Stanford dependency parser [7] to the sentence and examine all "paths" in the dependency tree containing the word of interest (WI). Depending on the position of the WI in the path, it is either shortened or augmented by another branch in the dependency tree. The resulting sentence snippet is then:

Do you mean 'orange' like "orange and purple crayon streaks"?

This approach does have some interesting failure modes – sometimes the question doesn't help as much as one might hope, especially if the context is shortened too much; e.g.,

SENTENCE = The orange prize for fiction celebrates 10 years with best of the best.
QUESTION = Do you mean 'orange' like 'orange prize'?

Unless you know that the Orange Prize is the name of a prize for novels that is sponsored by Orange S.A. it is hard to know that this means "Orange like the company" (indeed after inspection we can see that '`* prize`' is a good but not great spotter for organizations).

SENTENCE = "thunder iv gx cards are designed to work with apple's implementation of the nubus slot."
QUESTION = Do you mean apple like "work with apple's implementation"?

The above example shows how a sentence snippet clearly identifies the meaning of the word 'apple' referring to the computer company and not the multimedia company. More examples are in Table 2. In some cases no good example phrase was found in the ukWaC corpus.

Type	**Suggestion**
Fruit	Do you mean orange like "coffee or orange juice" ?
Fruit	Do you mean blackberry like "blackberry juice" ?
Fruit	Do you mean apple like "orange and apple juice" ?
Company	Do you mean blackberry like "explore the possibilities of blackberry beyond email" ?
Company	Do you mean orange like "orange prize" ?
Company	Do you mean apple like "works with apple's implementation" ?
Company	Do you mean sage like "register with sage" ?
Color	Do you mean orange like "orange and purple crayon streaks"?
Color	Do you mean apple like "blue and apple red"?
Color	Do you mean sage like "cut from colorbok sage stripe"?

Table 2. Examples of clarifying questions using phrases obtained from the corpus.

5 Conclusion

Finding the entity for an ambiguous mention can be challenging in cases where there is not enough context for automatic disambiguation. As knowledge bases become more complete and broadly applicable, this problem will only increase. In cases where systems are interacting with users, it may be advantageous to summarize the ambiguity in a question that present alternative entities in natural language for users to select.

We present three methods for implementing clarifying question generation to interact with users of these systems. It is clear there is no one method that works all the time, and that this task may best be done with an ensemble of methods. We hope that our discussion in this paper will inspire the creation of many more methods and improve the interaction of knowledge-based systems for complex tasks.

Rigorous evaluation of our preliminary work is the focus of our ongoing effort. One of the challenges is that the "goodness" of a clarifying question is an inherently subjective opinion. We are also investigating more sophisticated methods that may mitigate some of the limitations we discussed here. Another research topic is on extending our approaches to be able to determine clarifying questions for entities which are not members of any known type - in effect building a system to augment and create existing taxonomies and ontologies.

Acknowledgements

Authors listed in alphabetical order by last name. We would like to thank Alfredo Alba, Clemens Drews, Linda Kato, Chris Kau, Steve Welch and others that have helped in the development of Concept Expansion.

References

1. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In: Language Resources and Evaluation 43 (3). pp. 209–226 (2009), http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf
2. Coden, A., Gruhl, D., Lewis, N., Tanenblatt, M.A., Terdiman, J.: SPOT the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In: IEEE HISB 2012. pp. 33–39 (2012)
3. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: I-SEMANTICS. pp. 121–124 (2013)
4. De Boni, M., Manandhar, S.: An analysis of clarification dialogue for question answering. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. pp. 48–55. NAACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1073445.1073452
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal (2014)
6. Loos, B., Porzel, R.: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004, chap. Resolution of Lexical Ambiguities in Spoken Dialogue System (2004), http://aclweb.org/anthology/W04-2312
7. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (2006)

8. McCray, A., Aronson, A., Browne, A., Rindflesh, T., Razi, A., Srinivasann, S.: UMLS knowledge for biomedical language processing. Bull Med Libr Assoc.;81:184194 (1993)
9. Mendes, P.N., Jakob, M., Bizer, C.: DBpedia for NLP: A Multilingual Cross-domain Knowledge Base. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). pp. 23–25. Istanbul, Turkey (2012)
10. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: shedding light on the web of documents. In: Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011. pp. 1–8 (2011)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Qadir, A., Mendes, P.N., Gruhl, D., Lewis, N.: Semantic lexicon induction from twitter with pattern relatedness and flexible term length. In: AAAI 2015. pp. 2432–2439 (2015)
13. Savova, G.K., Coden, A.R., Sominsky, I.L., Johnson, R., Ogren, P.V., Groen, P.C.d., Chute, C.G.: Word sense disambiguation across two domains: Biomedical literature and clinical notes. Journal of Biomedical Informatics 41(6), 1088–1100 (2015/12/22 2008), http://dx.doi.org/10.1016/j.jbi.2008.02.003
14. Schlangen, D.: Causes and strategies for requesting clarification in dialogue. In: Proceedings of the 5th Workshop of the ACL SIG on Discourse and Dialogue (2004)
15. Stoyanchev, S., Liu, A., Hirschberg, J.: Towards Natural Clarification Questions in Dialogue Systems. In: AISB Symposium on "Questions, discourse and dialogue: 20 years after Making it Explicit," AISB-50. Goldsmiths, London, UK (2014)