Road Segment Identification in Natural Language Text

Ahmed Y. Tawfik and Lawrence Barsanti

School of Computer Science, University of Windsor
Windsor, Ontario, Canada
Email atawfik@uwindsor.ca, barsant@uwindsor.ca

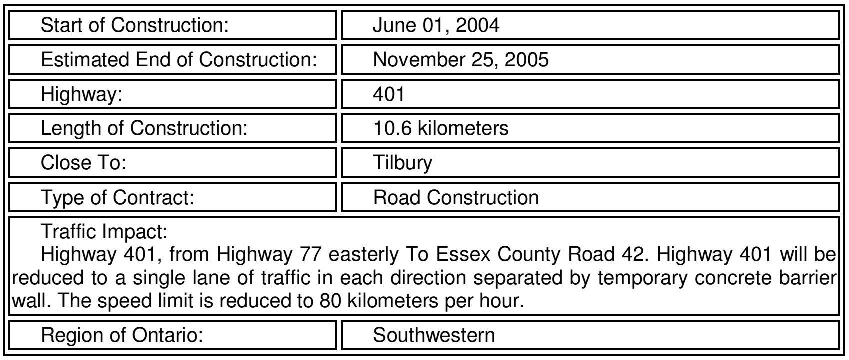
Abstract. This paper describes a technique to extract geographic location information from a natural language description of a location. The technique relies on a set of domain specific tags and a set of keywords. The tags are used to identify roads, intersections, and landmarks. Tag combinations are used to discover road segments. The technique is applied to understanding highway construction reports for the Canadian Province of Ontario.

1. Introduction

Location information has traditionally been expressed in two main forms: natural languages and maps. Maps represent a rich visual representation that captures a host of spatial relationships among collocated elements. Natural languages provide a set of focused abstractions of the spatial relationships represented in a map. In natural languages, the choice of the relevant abstraction is generally task dependent. Maps and natural language interfaces to geographic information systems continue to be complementary. For example, systems that generate driving directions like Yahoo! Maps, MapPoint and MapQuest [6] provide a linguistic description of a map. Coral [1] applies natural language generation techniques to make the linguistic description more natural. Understanding and visualizing textual geographic references on a map has attracted less attention as a research focus. By grounding named entities to spatial locations, a system can answer spatial queries [3]. A geo-parser combines data from multilingual gazetteer with natural language text and a geographic information system to produce a map highlighting the locations mentioned in the text [4].

The focus here is on defining a set of special purpose tags that are designed to understand urban location descriptions like driving directions that can be used in translating location information expressed in natural language to a segment or region on a map. The application that has motivated this work is building a system that determines the location of highway construction based on construction report summaries. These summaries include some structured fields (e.g. affected highway, closest city, length of construction) and a natural language description of the traffic impact. The traffic impact typically includes detailed location information. Figure 1 shows an example of highway construction summary for highway 401 in Ontario[1], Canada.

[1] From the Ontario highway construction reports available at http://www.mto.gov.on.ca/english/index.html



Start of Construction:	June 01, 2004
Estimated End of Construction:	November 25, 2005
Highway:	401
Length of Construction:	10.6 kilometers
Close To:	Tilbury
Type of Contract:	Road Construction
Traffic Impact: Highway 401, from Highway 77 easterly To Essex County Road 42. Highway 401 will be reduced to a single lane of traffic in each direction separated by temporary concrete barrier wall. The speed limit is reduced to 80 kilometers per hour.	
Region of Ontario:	Southwestern

Figure 1. Sample Construction Report Summary

Section 2 presents the knowledge representation that serves as a foundation to the work. Section 3 introduces the two level parsing technique used in the interpretation of some natural language location descriptions. Section 4 presents the results of analysing construction reports and Section 5 presents a brief conclusion.

2. Elements of the Knowledge Representation

Topological and metric spatial relationship expressed in natural language has to be interpreted before the location can be correctly determined. In general, we consider that we have linear entities and regions. A road is represented as a linear entity. Linear entities include highways, creeks, rivers, and boundary lines. Towns, cities, counties, and mountains are considered as regions. The intersections of two lines define a point. The intersection of a line and a region defines a line segment. Specifying a location relies on the identification of the relationship that holds between lines or between a line and region. Interpreting the natural language terms describing these relationships relies on the two-level part-of-speech-tagging described in the next section.

The knowledge representation is based the 9-intersection model [5]. According to this model, each spatial object divides the space into three components: the boundary of the object, the space internal to the object, and every thing else is external to the object. Therefore, a simple line (that has no self loops) has two boundary points, and a continuous sequence of internal points joining the two boundary points. Similarly, a region has a closed boundary, an internal area and an external area. For simplicity, we assume that the map is a 2D space.

Line-line Relations

Shariff et al. [5] identify 33 topological relationships that may hold between two lines. To simplify the representation, we omit self-similar (symmetric) relationships. A relationship is self-similar if its inverse has the same 9-intersection matrix as the original relationship The 33 relations include 11 self-similar relationships in addition to 11 relationships with 11 respective inverses. . For example, ***equal*** (LL22) and ***intersect*** (LL2) are self-similar relationships. However, ***contains*** (LL5) has an inverse (LL5^{-1}).

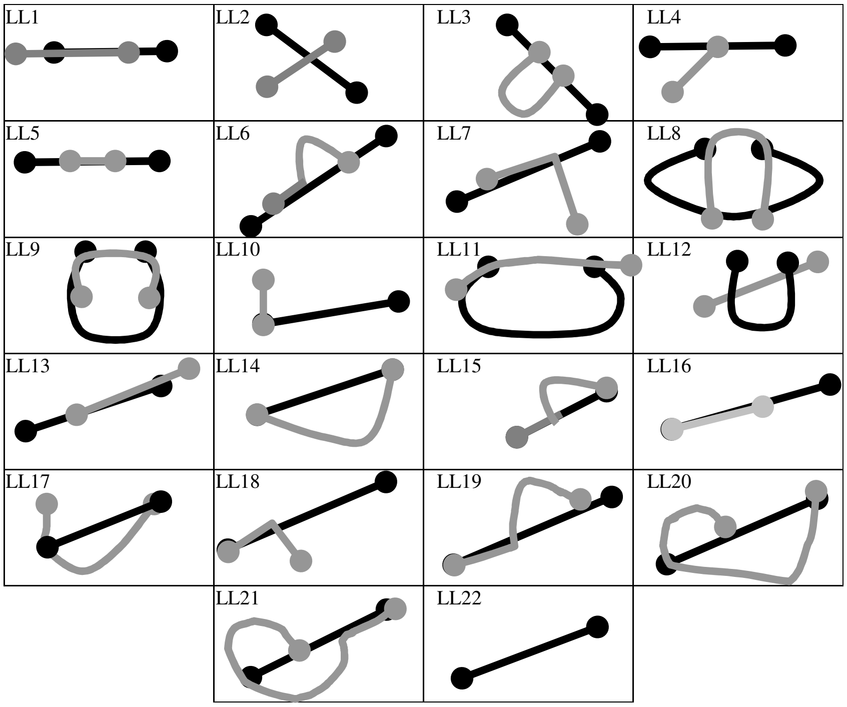


Figure 2. Line-Line Relationships

Figure 2 illustrates the topological relationships that may hold between two lines. It is possible to divide these topological relationships into relationships involving overlapping segments (LL5, LL6, LL7, LL11, LL12, LL15, LL16, LL18, LL19, LL21, and LL22) and others involving 0, 1 or 2 boundary points (LL1, LL2, LL3, LL4, LL8, LL9, LL10, LL13, LL14, LL17, and LL20).

3. Parsing Location Information

In order to understand location text, it is important to extract references to locations in the text. In general, this a difficult problem as special attention should be given to the use of prepositions (at, from, to, …etc.) and a great deal of disambiguation may be involved in distinguishing references to places from other proper nouns. A gazetteer is useful in distinguishing references to cities and towns from other proper nouns in the text. However, some commonsense knowledge is necessary to correctly parse "Mr.

England flew to India" and figure out that England refers to a person while India is a place. Fortunately, in our application, location information was easily identifiable and a rather limited amount of effort went into disambiguation. The tagging technique introduced here assigns special tags to spatial words and phrases indicating direction as in northerly, left, southbound, and Windsor-bound. Special tags are also assigned to words and phrases denoting proximity or distance like near, close, next to, or distance (like a mile). As some numbered highways also have a name (e.g. County Road 19 is also Manning road), a special tag (ALT_ROAD) is necessary. Table 1 lists the set of domain specific tags used here.

Table 1. Domain Specific Tags

Tag	Represents
INT_ID	Intersection Identifier
ROAD	Road / Highway
OFF	Offset/Proximity/Distance
DIR	Direction of Traffic (set of lanes)
NLM	Natural Landmark
MLM	Manmade Landmark
ALT_ROAD	Alternate Road Name

The assignment of these domain-specific tags is performed as a second level tagging after the text has been tagged using a standard part-of-speech tagger. Here, we use CLAWS [2] for first level tagging. CLAWS tag set includes locative tags NNL, NNL1 and NNL2). However, we found that the names of most roads and landmarks consist of a sequence of singular common nouns (NN1), singular proper nouns (NP1), numbers (MC), and in some cases title nouns (NNSB). Phrases that contain these tag sequences are of interest, we identify these phrases as potential name phrase (PNP). A PNP is defined as a sequence of one or more words whose tags are any combination of the NN1, NP1, and NNB tags; see Figure 3. Locative tags still play an important role in identifying locations. For example, both roads and natural landmarks can be found by searching for the sequence of tags PNP NNL1. It is the word represented by the NNL1 tag that distinguishes between them. That is why some of the tag sequences presented in Table 3 have keywords associated with them. Table 2 lists the keywords used in assigning spatial tags. Adjectives (JJ), prepositions (II), nouns of direction (ND), units of measurements (NNU), the preposition "for" (IF), and participle (or past) form of verbs (VVN/VVD) all proved useful in identifying road segments.

Text	Highway	77	,	From	Highway	77	Easterly	To	Essex	County	Road	42	.
Tags	NN1	MC	,	II	NN1	MC	JJ	II	NP1	NN1	NN1	MC	.
	PNP				PNP				PNP				

Figure 3. Example of Potential Name Phrase Tags

Notice that in Table 3 some of the patterns contain domain specific tags. For this reason the order in which tags are found is important; the following order is used: INT_ID, PNP, ROAD, ALT_ROAD, OFF, DIR, NLM, MLM.

Table 2. Keyword Lists

Feature	Keywords
Road Indicator	avenue, boulevard, parkway, way, expressway, drive, road
Numbered Roads	Highway, route, road
Natural landmark	river, creek, brook, lake, island, isle, islet, narrows, mountain, forest
Manmade landmark	bridge, span, overpass, underpass, tunnel, structure, culvert, skyway
Direction	Northbound, northward, southbound, southward, eastbound, eastward, westbound, westward
Destination	Bound
Intersection	Intersection, junction, crossroad, crossway, crossing, corner, interchange
Road type	regional, municipal, county
Directional Adjective	Northerly, southerly, easterly, westerly

Table 3. Patterns for detecting domain specific tags

Special Tag	Sequence	Keywords From Table 2	Example
INT_ID		Intersection	
ROAD	PNP MC	PNP ends with Numbered Rd	County Road 42
	PNP NNL1	NNL1 not a landmark	Queen Street
	PNP NNL1 MC	NNL1 not a landmark	County Road 121
	JJ PNP MC	JJ Road Type	Regional Road 3
	PNP	Starts with a numbered road or ends with a road indicator	Highway QEW Van Horne Ave.
OFF	ND1 IO		East of
	MC NNU1 ND1 IO		1 kilometer East of
	MC NNU2 ND1 IO		2 kms East of
	JJ MC NNU1	JJ directional adjective	Northerly 1.0 km
	JJ MC NNU2	JJ directional adjective	northerly 2.3 kms
	JJ IF MC NNU1	JJ directional adjective	Northerly for 1 km
	JJ IF MC NNU2	JJ directional adjective	Northerly for 2 kms
DIR	ROAD ANY	ANY is direction	Highway 401 westbound
	ANY ROAD	ANY is direction	Eastbound Highway QEW
	PNP VVD	VVD is destination	Toronto Bound
	ND1 VVN	VVN is destination	west bound
	ANY	ANY is direction	Westbound
	ROAD ND1		Highway 8 East
NLM	PNP	Last word natural landmark	Pike Creek
	PNP NNL1	NNL1 in natural landmark	
MLM	NLM ANY	Any in manmade landmark	Pike Creek Bridge
	PNP	Last word in manmade landmark	Thorold Tunnel
ALT_ROAD	(ROAD)		(Manning Road)

However, using only patterns does not provide good enough results. For instance some people may use short forms like “Take Highway 401 to Walker Road. Get off the 401 and go south on Walker.” In this sentence both “the 401” and “Walker” refer to roads, but would not be matched by any of the listed patterns. To handle this type of situation every time a road is found the rules in Table 4 are used to generate potential short form for the road. After all the patterns have been searched a second pass can find and tag these short forms.

Lastly, in some cases CLAWS tags a word like road as a noun (NN1) when it should be a locative noun (NNL1). That is why there are redundant rules like “proper noun followed by a number” (PNP MC) and “proper noun followed by a common noun and a number” (PNP NNL1 MC).

Identifying Road Segments

Usually, road segments are defined as either the stretch of road between two points (Figure 4a), or as a stretch of a given length starting at a given point (Figure 4b). For example, construction may affect a highway segment between two intersections or it may affect the area around a bridge or intersection. According to the 9 intersection model, a point is the intersection of two lines (LL2) or a segment and a line (LL4). A segment can also be defined in terms of the portion of a line intersecting a region. In this case, the points defining the ends of the segment are defined by the intersections of the region boundaries with the line.

Table 4. Rules for creating short forms

Pattern	Short Form
PNP MC	“the MC” and “the highway” (if PNP is the word highway)
PNP NNL1	PNP (i.e. road name without street, road, etc…)
PNP NNL1 MC	“the MC”
JJ PNP MC	“the MC”
PNP	“the PNP” without first word or PNP without last word

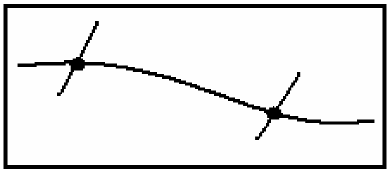


Figure 4a. Segment delimited by two points

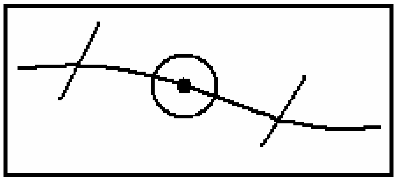


Figure 4b. Segment around a point

Table 5. Segment Identification Rewriting Rules

Type	Pattern	Rewrite As
Segment with two points	ROAD^1 ROAD^2 to ROAD^3	ROAD^1 & ROAD^2 to ROAD^1 & ROAD^3
	ROAD^1 ROAD^2 to NLM	ROAD^1 & ROAD^2 to ROAD^1 & NLM
	ROAD^1 NLM to ROAD^3	ROAD^1 & NLM to ROAD^1 & ROAD^2
	ROAD^1 NLM^1 to NLM^2	ROAD^1 & ROAD^2 to ROAD^1 & MLM
	ROAD^1 ROAD^2 to ROAD^3	ROAD^1 & MLM to ROAD^1 & ROAD^2
	ROAD^1 between ROAD^2 and ROAD^3	ROAD^1 & ROAD^2 to ROAD^1 & ROAD^3
	Note: if DIR was found using either the 1st or 2nd pattern in Table 3 it could substitute for one of the ROAD tags. i.e. ROAD westbound ROAD to MLM	
Segment with a single point	ROAD^1 at ROAD^2	ROAD^1 at ROAD^2 (no interjecting words)
	ROAD at NLM	ROAD at NLM (no interjecting words)
	ROAD at MLM	ROAD at MLM (no interjecting words)
	ROAD^1 OFF^1 ROAD^2 OFF^2	ROAD^1 at ROAD^2
	ROAD^1 ROAD^2	ROAD^1 at ROAD^2
	ROAD NLM	ROAD at NLM
	ROAD MLM	ROAD at MLM

Table 5 shows how the domain specific tags and CLAWS part-of-speech tags are used to identify road segments. The sequences used to identify road segments consist of tags and keywords. The tags and keywords in a pattern have to appear in the order laid out by the pattern; but they do not have to be consecutive. For example, the sentence in figure 5 matches the first pattern in Table 5 and produces the phrase "Highway 58 & Regional Road 3 to Highway 58 & Forks Road". The rewriting rules in Table 5 try to deal with implicit references to points and intersections. For example, in the phrase "Construction on Highway 22 from Howard Ave to Walker Rd." implicitly means that the construction starts at the intersection of Howard Ave. and Highway 22 and ends at the intersection of Walker Road and Highway 22.

<table>
<tr><th>Text</th><td colspan="7">Highway 58 , Regional Road 3 To South Of Forks Road .</td></tr>
<tr><th>Tags</th><td>ROAD</td><td>,</td><td>ROAD</td><td>II</td><td>OFF</td><td>ROAD</td><td>.</td></tr>
</table>

Figure 5. Example of Road Segment Identification

4. Implementation and Performance Results

The algorithm described in this paper has been implemented in Java. The implementation consists of two main components: a special purpose tagger and a road segment identification module. It uses a separate file for the keywords and patterns shown in Tables 2, 3, 4, and 5 to maintain modularity and make it easy to add new patterns or keywords. An input text is first submitted to CLAWS to get a part-of-speech (POS) tagged text. In the POS tagged sequence, keywords are identified and the sequence is matched to the patterns given in the keywords/patterns file to get a text tagged with our application-specific tags in addition the part-of-speech tags. If a word could have more than one POS tag, the tags are considered in the order of their likelihood until a matching pattern is found. If a word sequence matches multiple patterns, (e.g. PNP NNL1 and PNP NNL1 MC), the longest sequence, that overlaps the shorter one, is used. The road segment identification module applies the rewriting rules in Table 5 and generates a list of road segments. The algorithm looks for road segments delimited by two points first before trying to identify road segments that are near a point. The following example illustrates the outputs from the tagger and the segment identifier.

Input	*Hwy*	*401*	,	10	Kms	east	of	Interchange	Number	661	*At*	the	*Donovan*	*Creek*	*Bridge*
POS	NN1	MC	,	MC	NNU2	ND1	IO	NN1	NN1	MC	II	AT	NP1	NN1	NN1
Domain Tag	ROAD			OFF				INT_ID					MLM		
Segment	(ROAD AT MLM) OFF: Hwy 401 at Donovan Creek Bridge, 10 Kms East														

To test the performance of the tagger, we used 25 construction reports from Ontario road construction web site. Each construction report was manually tagged using the domain specific tag set, and then the reports were tagged by the tagger. The tags found by the domain tagger were then checked for incorrect tagging. The results are summarized in Table 6.

From the results it clear that the patterns for both natural and manmade landmarks contributed a large number of errors. There are too many false positives for natural landmarks. A more semantic approach to finding natural landmarks should reduce the number of false positives because the false positives all occurred when the tagger failed to find either a manmade landmark (i.e. the rule NLM ANY failed because of the keyword list) or road. As for manmade landmarks, a wider range of patterns and better generalization would probably improve their results as the test set simply contained keywords that were not identified in the initial analysis.

To evaluate the performance of the road segment identification, twenty-three of the twenty-five construction reports used for domain tagging were read by a person who looked for occurrences of road segments as explained in Figure 4. The tagged reports were processed by the system and the strings it produced were reviewed for correctness and counted. Additionally, all of the misses and partially correct (e.g. near A instead of between A & B) results were examined to determine if the problem was a consequence of the domain tagging. Table 7 presents a summary of the results.

We found that inaccuracies in the tagging seriously degraded the quality of the information extracted. A second look at the tagging results revealed that even though the domain tagger has an average accuracy of 76.7% only 12 of the 25 reports (48%) were tagged flawlessly. In fact, every other report had at least one error, which could easily throw off the segment identification.

Table 6. Tagging Results

Tag	Actual	Total Found	False Positives	Percent Correct
INT_ID	12	12	0	100
ROAD	63	50	2	76.2
OFF	28	26	1	89.3
DIR	17	13	0	76.5
NLM	1	4	4	0
MLM	14	7	1	42.9
ALT_ROAD	2	1	0	50
Weighted Average				76.7

Table 7. Road segment identification results

Type	Actual	Correctly Identified with automatic tagging	Correctly Identified with manual tagging
Two Points	10	3	9
Single Point	19	12	16
Other	1	0	0
Overall	30	15	25

We then added a CITY tag that matches the city name from the structured information associated with the construction report. The CITY tag is useful to avoid interpreting a phrase like “Riverside Drive, Windsor” as an intersection.

Surprisingly, in our tests of the road segment identification algorithm, no false positives were produced; however in some cases weaker, but correct information was found (e.g. near A instead of between A & B). This is surprising because the last three patterns in Table 5 are rather lenient. Also, note that information such as direction and offset could also be leveraged in order to improve the extracted information. For instance, the pattern ROAD1 OFF1 ROAD2 OFF2, occurred in four reports and was defined as a region around an intersection, but it would have been better defined as a road segment delimited by two points.

5. Discussion and Conclusions

Determining the location of the highway construction described in Figure 1 requires parsing the text to extract the starting landmark or intersection (Highway 401 and Highway 77) and the ending landmark or intersection (Highway 401 and Essex County Road 42). Using a gazetteer as a dictionary to look up road names, landmarks, and populated places should improve the performance [4]. However, in many cases the construction zone is bound by harder to define landmarks. Consider the following examples:

- "Highway 35, Victoria/Haliburton Boundary Northerly for 8.1 kilometres to 0.3 kilometres north of Miners Bay."
- "Highway 21, from the north limits of Goderich northerly for 2.0 kilometres and Straughn's Creek Culvert 8 kilometres south of the town of Goderich."
- "Highway 401, from 2.55 kilometres west of Boundary Road, easterly to 0.75 kilometres east of Boundary Road and westbound lanes 0.5 kilometres east of Brookdale Avenue, easterly for 0.5 kilometres."

In the first example, the construction zone is defined in terms of the highway intersection with the boundary between two regions. In the second example, the construction zone is apparently discontinuous as it spans two kilometres from the north limits of a town and 8 kilometres from a small creek to the south of the same town. In the third example, the word "Boundary" is the name of a road not a region boundary as in the first example.

This work has provided a technique to identify road segments that are of interest for some reason (in our case, they were affected by construction). The technique can be useful in other applications like understanding driving directions. The results reported here are for a relatively small test corpus obtained from a single source and may not be statistically significant. However, these results highlight some of the strengths and limitations of the proposed approach.

Acknowledgements

This work is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC), Canada.

References

1. Dale, R., Geldof, S., and Prost, J.-P.: CORAL: Using Natural Language Generation for Navigational Assistance", Twenty-sixth Australasian Computer Science Conference (ACSC2003), Adelaide, South Australia, (2003).
2. Garside, R., and Smith, N. : A Hybrid Grammatical Tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, (1997) 102-121.
3. Leinder, J., Sinclair, G., Webber, B.: Grounding Spatial Named Entities for Information Extraction and Question Answering. Workshop on the Analysis of Geographic References at NAACL-HLT 2003 Conference, Edmonton, Alberta, Canada, (2003).
4. Poluiquen, B., Steinberger, R., Ignat, C. and De Groeve, T.: Geographical Information Recognition and Visualization in Texts Written in Various Languages. The 2004 ACM Symposium on Applied Computing (SAC'04), Nicosia, Cyprus, (2004) 1051-1058.
5. Shariff, R.B.M., Egenhofer, M.J., Mark, D.M.: Natural-Language Spatial Relations between Linear and Areal Objects: The Topology and Metric of English-Language Terms. International Journal of Geographical Information Science, Vol. 12 No. 3, (1998) 215-246.
6. Forth, S.: Online Mapping Services Guide the Way. Plugged In, 16 :11 (2005) 48-50.