Identifying Collisions in NTSB Accident Summary Reports

Torbjörn Ekman
Anders Nilsson
Dept. of Computer Science, Lund University
Box 118
S-221 00 SWEDEN
Tel: +46 46 222 01 85
Fax: +46 46 13 10 21
email: {torbjorn|andersn}@cs.lth.se

June 13, 2002

Abstract

Insurance companies are often faced with the task of analyzing car accident reports in order to try to find out how the accident took place, and who to blame for it. The analysis could in many cases become much easier if the accident could be automatically visualized by extracting information from the accident report.

This paper describes the collision detection functionality of the text processing system that is used with CarSim, a car accident visualization system. Using Link Grammar and regular expression pattern matching, we have correctly extracted 60.5% of the collisions found in 30 authentic accident reports.

1 Introduction

CarSim [DELN01] is system for visualizing car accidents from natural language descriptions developed at the University of Caen, France. The aim of the CarSim system is to help users to better understand how a car accident took place by visualizing it from the description given by the participants in natural language. The visualization process is performed in two steps. First comes an information extraction step where the natural language texts are analyzed, relevant information is extracted[1] and templates are filled with the extracted information. The second step reads templates and constructs the actual visualization.

As part of further improvements in CarSim, a new information extraction system for English texts is being developed at the Department of Computer Science, Lund University. This paper describes how collisions are detected, and how to find extract collision verbs together with their subject/object pairs. In some cases it is also possible to resolve co-references in the analyzed text.

[1] Not only the collision as such, but also other vehicles, trees, rocks, road signs etc.

1.1 NTSB Accident Reports

The National Transportation Safety Board (NTSB) [NTS] is an independent federal agency in the United States investigating every civil aviation accident in the US, and also significant accidents in other modes of transportation, such as highways and railroads. The aim is to understand why the accidents occurred and issue safety recommendations to prevent similar future accidents.

The NTSB publishes accident reports, and for many of them, shorter summaries, describing investigated accidents. The accident reports summaries, were used to test the implemented collision detector.

1.2 LinkGrammar

In the beginning of the 90's, Sleator and Temperly [ST] defined a new formal grammatical system they called a *link grammar*, related to the larger class of dependency grammars. The idea behind *link grammar* was to let the words of a sentence create a connected graph with no crossing arcs. The arcs connecting words are called links. These links describe the syntactic and semantic relationship between the connected words, i.e. connect a determiner to a noun or a noun phrase to a verb.

1.3 WordNet

WordNet [Fel] is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet was developed by the Cognitive Science Laboratory at Princeton University.

WordNet is in this project used for extracting all possible collision verbs, see section A, that could occur in the accident reports.

WordNet is currently being adapted to other languages, notably Swedish at the Linguistics Department at Lund University.

2 Subject and Object Detection in Collision Context

The subject and object detection is divided into three stages. First, a shallow analysis locates sentences that may describe collisions. That analysis uses regular expressions to find the texts containing collision verbs. These sentences are passed on to Link Grammar which performs a much more in depth analysis. The subject and object are finally detected in the link representation of a sentence by using a set of patterns describing how to locate those entities starting out with a collision verb.

2.1 Detecting Sentence Candidates

Using Link Grammar to analyze sentences is a quite costly operation. Although the accident reports are fairly short it is not very practical to analyze all sen-

tences when detecting collisions. Therefore an initial pass to collect a set of candidate sentences suitable for deeper analysis is performed.

The accident reports are first tokenized to form a set of sentences. Each sentence is then matched against a regular expression accepting sentences including collision verbs. Each accepted sentence is passed on to Link Grammar, a much deeper analysis engine.

The regular expression used is built to accept a sentence including any collision verb from a list of verbs. A complete list of collision verbs, extracted from WordNet, is available in Appendix A.

2.2 Linkage Patterns

The result from the Link Grammar analysis is a graph where a word has a number of arcs connected to it. Each arc has a direction, left or right, as well as a semantic meaning. The transition from one word to any connected word can be expressed as a series of link and word pairs, where the link has a direction and a type, and the word has a name and a type. We call such an ordered list of pairs a Linkage Pattern.

Our analysis engine processes a Linkage Pattern where each pair is expressed by a regular expression. This way a large set of link-word pairs may be expressed through a compact notation. The engine matches the regular expressions and returns the set of words that is reached by following the entire pattern. The regular expressions is used by the system to locate subject, object, and also to resolve a class of co-references.

2.3 Extracting Subject and Object

Starting out from the collision verb a set of Linkage Patters are matched against the Link Grammar representation of the candidate sentences. These Linkage Patters describe how to reach the subject and object when starting on the collision verb. In its simplest form, a pattern may find the subject by following a Subject link to the left from the collision word.

2.4 Handling Co-references

The Linkage Pattern technique can also be used to resolve co-references. In this case, a pattern describes how to reach the subject starting with a pronoun. Because Link Grammar only processes a single sentence, the subject and pronoun must be in the same sentence for this technique to be successful.

3 Result Analysis

The collision detector was run on 30 NTSB accident report summaries containing a total of 43 collisions. The result is shown in table 1, and the individual results for each collision report with a short comment can be found in appendix B.

Analyzing the erroneous and incomplete collision detections confirms the suspicion that the regular expression pattern matching works correctly, but that

Correct detections	26
Erroneous detections	5
Incomplete detections	12
Total no. of collisions	43
Hit Ratio	60.5%

Table 1: Results from NTSB accident reports.

the linkages returned from Link Grammar for these sentences are either incomplete or incorrect. The accident report summaries are written in bureaucratic American English with sometimes very long sentences with many subordinate clauses which make it very hard for Link Grammar to find complete linkages.

4 Related Work

Mokhtar and Chanod [AMC] uses another approach for extracting subject-object relationships. They use an incremental finite-state parser to annotate the input string with syntactic markings, and them pass the annotated string through sequence of finite-state transducers. Their results are much better than our experience with Link Grammar with a precision $\approx 90\%$ for both subjects and objects.

Ferro et al. [FVY99] uses a trainable error-driven approach to find grammatical relationships. The achieved precision running on a test set is $\approx 77\%$.

Brants et al. [BSK97] have a slightly different goal with their research. They have developed an interactive semi-automatic tool for constructing treebanks. Their results are very good, $> 90\%$, but then one has to have in mind that the only automatic annotation is to assign grammatical function and/or phrase category.

5 Conclusions and Future Work

We have shown that using Link Grammar together with regular expression pattern matching is a plausible technique for extracting collisions from, also grammatically complicated, texts. Achieving a hit ratio of 60.5% on the NTSB report summaries is a good result considering both the simplicity of the collision detector and the grammatical complexity of the tested sentences.

Since most of the collision detection problems originate in incomplete and/or incorrect linkages returned from Link Grammar, this is the obvious first choice for possible improvements. Adding domain-specific knowledge to Link Grammar so that it would concentrate on linking collision verbs with their subject and object instead of trying to create a complete linkage over the complete sentence would significantly enhance the hit ratio of the collision detector.

Other improvements to the collision detector to further enhance the hit ration somewhat include extending the regular expression patterns to match even more complicated linkages than is possible in the current implementation.

References

[AMC] Salah Aït-Mokhtar and Jean-Pierre Chanod. Subject and object dependency extraction using finite-state transducers. Rank Xerox Research Centre, Meylan, France.

[BSK97] Thorsten Brants, Wojciech Skut, and Brigitte Krenn. Tagging grammatical functions. In *Proceedings of EMNLP-2*, July 1997.

[DELN01] Sylvain Dupuy, Arjan Egges, Vincent Legendre, and Pierre Nugues. Generating a 3d simulation of a car accident from a written description in natural language: The CarSim system. In *Proceedings of The Workshop on Temporal and Spatial Information Processing*, pages 1–8. ACL, July 2001.

[Fel] Christiane Fellbaum. English verbs as a semantic net. *http://www.cogsci.princeton.edu/ wn/*.

[FVY99] Lisa Ferro, Marc Vilain, and Alexander Yeh. Learning transformation rules to find grammatical relations. In *Computational Natural Language Learning*, pages 43–52. ACL, June 1999.

[NTS] The national traffic safety board. *http://www.ntsb.gov*.

[ST] Daniel D. Sleator and Davy Temperly. Parsing english with a link grammar. *http://www.link.cs.cmu.edu/link/*.

A Collision Verbs

This list of collision verbs in collected from WordNet. The WordNet Browser was used to find synonyms to the verb *strike* in its collide sense. That procedure was recursively repeated to find all appropriate collision verbs.

collide, clash To crash together with violent impact. *The cars collided. Two meteors clashed.*

crash, ram To undergo damage or destruction on impact. *The plane crashed into the ocean. The car crashed into the lamp post.*

hit, strike, impinge on, run into, collide with To hit against or come into sudden contact with something. *The car hit a tree. He struck the table with his elbow.*

rear-end To collide with the rear end of something. *The car rear-ended me.*

broadside To collide with the broad side of something. *Her car broad-sided mine.*

bump, knock To knock against something with force or violence. *My car bumped into the tree.*

run into, bump into, jar against, butt against, knock against To collide violently with an obstacle. *I ran into the telephone pole.*

B Tested Sentences

HAR0001 As the bus approached milepost (MP) 184.9, it traveled off the right side of the roadway into an "emergency parking area," where *it* **struck** *the back of a parked tractor-semitrailer*, which was pushed forward and struck the left side of another parked tractor-semitrailer.

Result:

1. it struck the back of a parked tractor-semitrailer
2. a parked tractor-semitrailer struck the left side of another parked tractor-semitrailer

Comment Both collisions correctly found. Co-reference of collision 1 not resolved.

HAR0002 As *the bus* approached the intersection, *it* failed to stop as required and **was struck by** *the dump truck*.

Result:

1. the bus was struck by the dump truck

Comment Collision correctly found.

HAR0101 *The bus* continued on the side slope, **struck** *the terminal end of a guardrail*, traveled through a chain-link fence, vaulted over a paved golf cart path, **collided with** *the far side of a dirt embankment*, and then bounced and slid forward upright to its final resting position.

Result:

1. *Collision not found*
2. a paved golf cart path collided with the far side of a dirt embankment

Comment Subject and object not founf at all in collision 1, and incorrect subject in collision 2. Both due to incorrect LinkGrammar linkages.

HAR0102 About 90 seconds later, *northbound Metrolink commuter train 901*, operated by the Southern California Regional Rail Authority, **collided** with *the vehicle*.

Result:

1. 901 collided with the vehicle

Comment Collision found, but subject not fully resolved due to incomplete linkage.

HAR0103 Abstract: On March 28, 2000, about 6:40 a.m. (sunrise was at 6:33 a.m.), *a CSX Transportation, Inc., freight train* traveling 51 mph **struck** *the passenger side of a Murray County, Georgia, School District school bus* at a railroad/highway grade crossing near Conasauga, Tennessee.

Result:

1. train traveling struck the passenger side of a County

Comment: Collision correctly found, though object is not completely resolved. Works with proper name substitution.

HAR9001 Comment: No collisions in this report.

HAR9002 About 7:34 a.m ., central daylight time, on Thursday, September 21, 1989, *a westbound school bus* with 81 students operated by the Mission Consolidated Independent School District, Mission, Texas, and *a northbound delivery truck* operated by the Valley Coca-Cola Bottling Company, McAllen, Texas, **collided** at Bryan Road and Farm to Market Road Number 676 (FM 676) in Alton, Texas.

Result:

1. Company collided Number

Comment: Incorrect subject and object due to erroneous linkage.

HAR9003 Comment: No collisions in this report.

HAR9101 About 5:40 p.m. on July 26,1990, *a truck* operated by Double B Auto Sales, Inc., transporting eight automobiles entered a highway work zone near Sutton, West Virginia, on northbound Interstate Highway 79 and **struck** *the rear of a utility trailer* being towed by a Dodge Aspen.

The Aspen then **struck** *the rear of a Plymouth Colt*, and *the Double B truck* and *the two automobiles* traveled into the closed right lane and **collided with** *three West Virginia Department of Transportation (WVDOT) maintenance vehicles.*

Result:

1. 79 struck the rear of a utility trailer
2. the Aspen struck the rear of a Colt
3. the B truck collided with three

Comment: Incorrect subject in collision 1. Collision 2 is correct. In collision 3 there are two subjects of which only the first is found.The incorrect object in collision 3 is correct when using proper name substitution.

HAR9201 Comment: No collisions in this report.

HAR9202 About 9: 10 a.m. on December 11, 1990, *a tractor-semitrailer* in the southbound lanes of 1-75 near Calhoun, Tennessee, **struck** *the rear of another tractor-semitrailer* that had slowed because of fog.

After the initial collision, *an automobile* **struck** *the rear of the second truck* and was in turn **struck** in the rear **by** *another tractor-semitrailer.*

Meanwhile, in the northbound lanes of 1-75, *an automobile* **struck** *the rear of another automobile* that had slowed because of fog.

Result:

1. *Collision not found.*
2. an automobile struck the rear of the truck

3. *Collision not found.*
4. an automobile struck the rear of another automobile

Comment: Collision 1 is found when using proper name substitution. Collision 3 is represented by a complicated linkage which is not properly dealt with. Collision 2 and 4 are correctly found though the object of collision 2 is not completely resolved.

HAR9301 During the descent, *the bus* increased speed, left the road, plunged down an embankment, and **collided with** *several large boulders*.

Result:

1. the bus collided with several large boulders

Comment: Correctly found collision.

HAR9302 *The bus* **struck** *a car*, overturned on its right side, slid and spun on its side, uprighted, and **struck** *another car* before coming to rest.

Result:

1. the bus struck a car
2. the bus struck another car

Comment: Both collisions correctly found.

HAR9401 About 3:13 p.m., Wednesday, March 17, 1993, *an Amerada Hess (Hess) tractor-semitrailer* hauling gasoline **was struck by** *National Railroad Passenger Corporation (Amtrak) train 91*.

The truck, which was loaded with 8,500 gallons of gasoline, was punctured when *it* **was struck**.

Result:

1. an Amerada tractor-semitrailer struck by Corporation
2. *Collision not found.*

Comment: Incorrect linkage in Collision 1. Collision 2 is no collision and is therefore not found though a collision verb occurs.

HAR9402 On May 19, 1993, at 1:35 a.m., while traveling south on Interstate 65 near Evergreen, Alabama, *a tractor* with bulk-cement-tank semitrailer left the paved road, traveled along the embankment, overran a guardrail, and **collided with** *a supporting bridge column* of the County Road 22 overpass.

An automobile and *a tractor-semitrailer*, also southbound, then **collided with** *the collapsed bridge spans*.

Contributing to the severity of the accident was the collapse of the bridge, after *the semitrailer* **collided with** and demolished *the north column*, that was a combined result of the nonredundant bridge design, the close proximity of the column bent to the road, and the lack of protection for the column bent from high-speed heavy-vehicle collision.

Result:

1. a tractor collided with a supporting bridge column of the County 22 overpass
2. an automobile collided with the collapsed bridge spans
3. *Collision not found.*
4. *Collision not found.*

Comment: Incorrect linkages in collisions 3 and 4.

HAR9403 About 3:30 p.m. CDT on May 28, 1993, *the towboat CHRIS*, pushing the empty hopper barge DM 3021, **collided with** *a support pier of the eastern span of the Judge William Seeber Bridge* in New Orleans, Louisiana.

Result:

1. CHRIS collided with a support pier of the Bridge

Comment: Collision found correctly.

HAR9501 Seconds later, *National Railroad Passenger Corporation (Amtrak) train number 88*, the Silver Meteor, carrying 89 passengers, **struck** *the side of the cargo deck and the turbine.*

Result:

1. Meteor struck the side of the cargo deck

Comment: Collision found, but the subject can not be resolved due to erroneous linkages.

HAR9502 *The truck* drifted across the left lane onto the left shoulder and **struck** *the guardrail*; *the tank* **hit** *a column of the Grant Avenue overpass.*

Result:

1. the truck struck the guardrail
2. the tank hit a column of the Avenue overpass

Comment: Both collisions found and resolved.

HAR9503 As *the lead vehicle* reportedly slowed from 65 miles per hour (mph) to between 35 and 40 mph, *it* **was struck** *in the rear.*

Subsequent collisions occurred as *vehicles* **drove into** *the wreckage area* at speeds varying from 15 to 60 mph.

Result:

1. the lead vehicle was struck the rear
2. *Collision not found.*

Comment: Collision 1 found and almost completely resolved. Collision 2 could not be found due to complicated (erroneous?) linkages.

HAR9601 About 35 minutes later, *the truck* **was struck by** *southbound Amtrak train No. 81* en route from New York City to Tampa, Florida.

Result:

1. the truck was struck by southbound Amtrak

Comment: Collision found, object could not be fully resolved.

HAR9602
Comment: No collisions in this report.

HAR9701S Abstract: On November 26, 1996, *a utility truck* **collided with** and fatally injured *a 10-year-old student* near Cosmopolis, Washington.

Result:

1. *Collision not found.*

Comment: Incorrect linkage.

HAR9702S Abstract: On April 25, 1996, *a truck* with a concrete mixer body, unable to stop, proceeded through an intersection and **collided with** and overrode *a passenger car* near Plymouth Meeting, Pennsylvania.

Result:

1. *Collision not found.*

Comment: Incorrect linkage.

HAR9801S Abstract: On June 11, 1997, *a transit bus* **collided with** *seven pedestrians* at a "park and ride" transit facility in Normandy, Missouri.

Result:

1. a transit bus collided with seven pedestrians

Comment: Correct.

HAR9801 *A flatbed truck* loaded with lumber, operated by McFaul Transport, Inc., that was traveling southbound on U.S. Route 41 **collided with** *the doubles truck*, lost control, and crossed over the median into the northbound lanes.

A northbound passenger van with nine adult occupants **struck** and underrode *the right front side of the flatbed truck* at the landing gear.

A refrigerator truck loaded with produce, operated by Glandt/Dahlke, Inc., *that* was also traveling northbound, **struck** *the right rear side of the flatbed truck.*

Result:

1. 41 collided with the doubles truck
2. *Collision not found.*
3. a refrigerator truck struck the right rear side of the flatbed truck

Comment: Collision 2 not found because of erroneous linkage.

HAR9802S Abstract: On October 9, 1997, about 12:10 a.m., *a truck tractor* pulling a cargo tank semitrailer was going under an overpass of the New York State Thruway when *it* **was struck by** *a sedan*.

The car **hit** *the right side of the cargo tank* in the area of the tank's external loading/unloading lines, releasing the gasoline they contained.

Result:

1. it was struck by a sedan
2. the car hit the right side of the cargo tank

Comment: Coreference in collision 1 not found, otherwise correct.

HZM9101 *The vehicle* overturned onto its side and **struck** *the embankment of a drainage ditch* located in a dirt field beside the road.

Result:

1. the vehicle struck the embankment of a drainage ditch

Comment: Correct.

HZM9901

Comment No collisions in this report.

RAR0201 Executive Summary: About 9:47 p.m. on March 15, 1999, *National Railroad Passenger Corporation (Amtrak) train 59*, with 207 passengers and 21 Amtrak or other railroad employees on board and operating on Illinois Central Railroad (IC) main line tracks, **struck** and destroyed *the loaded trailer of a tractor-semitrailer combination* that was traversing the McKnight Road grade crossing in Bourbonnais, Illinois.

The derailed Amtrak cars **struck** *2 of 10 freight cars* that were standing on an adjacent siding.

Result:

1. *Collision not found*
2. the derailed Amtrak cars struck 2 of 10 freight cars

Comment: Collision 1 not found because of erroneous linkage.

RHR9001 About 9:38 a.m., Pacific standard time, on December 19, 1989, *National Railroad Passenger Corporation (Amtrak) passenger train 708*, consisting of one locomotive unit and five passenger cars, **struck** *a TAB Warehouse & Distribution Company tractor semitrailer* in a dense fog at a highway grade crossing near Stockton, California.

Result:

1. *Collision not found.*

Comment: Incorrect linkage.