Journal of Siberian Federal University. Humanities & Social Sciences 1 (2008) 15-21
~ ~ ~

УДК 410

Natural Language and Computational Linguistics at the University of Sussex

John A. Carroll*
Department of Informatics, University of Sussex
Falmer Brighton BN1 9QJ, UK [1]

Received 1.09.2007, received in revised form 1.12.2007, accepted 15.01.2008

In this project we develop new ways of estimating the frequency distributions of the senses of words from raw (unannotated) text. We are starting to exploit these distributions to implement WSD systems which do not rely on the availability of hand-labelled resources. One way is to adopt the meaning distinctions set out in a particular dictionary or thesaurus and then to relate these distinctions to measures of similarity that we compute automatically between the occurrences of words in a text. Another way in which we use measures of word similarity is to automatically augment existing specialised thesauruses with new terms.

Keywords: computer processing of human language, natural language processing, COGENT, DELPHIN, RASP, SPARKLE, LEXSYS.

The Natural Language and Computational Linguistics group in the Department of Informatics at the University of Sussex is one of the largest groups in the UK of researchers focusing on computer processing of human language – also known as natural language processing (NLP). The group was formed over twenty years ago, and currently consists of around fifteen faculty, doctoral and postdoctoral researchers, headed by two professors. The group has a high profile internationally for research in both basic technologies for NLP, and also language-based computer applications.

Recent areas of research – and some national and European collaborative projects that have funded this research – are summarised below.

1. Basic Language Processing Technology

1.1. Word Meanings

-- Ranking Word Senses for Disambiguation

When faced with the question "Which plants thrive in chalky soil?" humans have no trouble understanding that the plants are floral rather than industrial. Furthermore, humans recognise that the answers "Sweetcorn and cabbage family vegetables do well on chalky soil", "Sweetcorn and cabbage grow well on chalky ground", and "Maize and cabbage-like vegetables grow well on chalky soil" are all paraphrases and mean more or less the same thing. Semantic interpretation and disambiguation is performed effortlessly by humans but poses great difficulties to computer-based applications that extract, filter and manipulate information from textual data. With

* E-mail address: J.A.Carroll@sussex.ac.uk

the rapidly growing amounts of text being stored by businesses and available over the Internet, such applications become increasingly important and the development of improved methods for identifying the intended meaning of words (word senses) will be a key technology.

The most accurate techniques for word sense disambiguation (WSD) to date are those trained on text in which each word has been manually annotated with its intended sense. A major shortcoming of these methods, though, is that accuracy is strongly correlated with the quantity of training data available, and this is in short supply because its production is very labour-intensive. For many words the distribution of their senses is highly skewed and WSD systems work best when they take the most frequent sense into account. However, the most frequent sense of a word is often not known, particularly in domains (subject areas) in which no text has ever been manually annotated.

In this project we are developing novel ways of estimating the frequency distributions of senses of words from raw (unannotated) text. We are starting to exploit these distributions to implement WSD systems which do not rely on the availability of hand-labelled resources.

-- Word Similarity

One way in which we are approaching the phenomenon of word meaning computationally is to adopt the meaning distinctions set out in a particular dictionary or thesaurus, and then relate these distinctions to measures of similarity that we compute automatically between occurrences of words in text.

Pioneering research at Sussex has comprehensively investigated a range of different measures of word similarity: these are based on various ways of comparing vectors of frequency counts, where elements of the vectors encode specific, pre-determined features of the context of occurrence of a word. Context features might be (for a noun), whether it is the object of a particular verb, whether it is modified by a particular adjective, etc. For example, the words “star” and “planet” often occur as the object of “view” and modified by “bright”. But “star” is also similar to “actress” (since they are both often modified by “famous”) although “planet” is not. These sorts of contextual differences can help us define ways of automatically teasing apart word meanings.

Another way in which we are using measures of word similarity is to automatically augment existing specialised thesauruses with new terms. One area in which this is especially important is in medical research, in which new types of diseases, pathogens and other micro-organisms, and drugs and treatments are being discovered at an ever increasing rate. Researchers in this area need to be able to keep track of new terminology and easily find out how it relates to existing terms. To tackle this problem, we are currently investigating a novel technique for automatically enlarging an existing medical thesaurus with new terms appearing in published scientific papers.

-- MEANING: Developing Multilingual Web-scale Language Technologies

Most computer applications that use the World Wide Web as a knowledge source process the information available only at the word level, making no attempt to go deeper and deal with the words’ meanings. The MEANING project was concerned with automatically collecting and analysing language data on a large scale, building (for several European languages) comprehensive knowledge bases about word meanings and their inter-relationships, and developing techniques for computing with word meanings. These knowledge bases and techniques can facilitate development of concept-based - rather than word-based - open domain Internet applications (such as Question/ Answering, Cross Lingual Information Retrieval,

Summarisation, Text Categorisation, Information Extraction, Machine Translation, etc.).

The main organisational structure for language data in the MEANING project was EuroWordNet, a large multilingual thesaurus covering several European languages, and representing translational correspondences between the languages through an `Interlingual Index'. Using this index, information about concepts acquired for one language can be ported across to all the other languages in the thesaurus. So, for example, certain types of concept-level information might be more easily or accurately found by processing English text (since there is a more of it available than other European languages), and then it could be distributed to the other languages.

Towards the end of the project, we started to treat the Web as a (huge) corpus to learn information from, since even the largest conventional text corpora available are not large enough to be able to acquire reliable information in sufficient detail about language behaviour, and most European languages do not have large or diverse enough corpora available.

1.2. Text Generation

-- COGENT: Controlled Generation of Text

Natural Language Generation (NLG) technology has reached a level of maturity where applied systems exist in a range of specialised real-world domains (such as weather bulletins, software documentation, health and legal advice and stock market movements). However, developing such systems currently involves hand-crafting and special-purpose tuning by NLG experts, which is non-portable, non-scaleable, time-consuming and expensive. Wider deployment of language generation requires more generally applicable and reusable NLG components based on wide-coverage grammars, but at present, effective techniques for such wide-coverage generation are not well understood.

In this project we are investigating systematically the characteristics of wide-coverage generation and developing reflective techniques for controlling it effectively. As well as furthering our understanding of wide-coverage generation, the project is delivering a substantial and novel resource to support future research in this area, and practical implementations of wide-coverage controllable generators.

-- The DELPH-IN (Deep Linguistic Processing with HPSG) Collaboration

Related to our work in the COGENT project, we have been collaborating with colleagues in Norway to implement algorithms for generating text using large, detailed grammars written by linguists. The main problem with using such grammars is efficiency of processing; we have devised a number of new approaches for dealing with the complexity of such grammars. The resulting system forms part of an automatic translation demonstrator from Norwegian to English in the domain of tourist information.

1.3 Text Analysis

-- Robust Accurate Statistical Parsing: RASP

Over the past few years we have been developing a robust, domain-independent parsing system for English, called RASP. The system takes text as input, and produces as output a set of relations encoding the grammatical dependencies between the words in each sentence. The system uses a combination of symbolic information (a hand-written grammar) and statistical information (probabilities associated with parts of the grammar that indicate what types of grammar structures are more likely than others, in order to deal with ambiguities). The RASP system is being used

by a large number of research groups worldwide within language processing applications.

The system forms the focus of further research at Sussex on:

– the development of fully domain-independent automated training regimes, allowing the rapid construction of an accurate parser for specific domain-dependent applications;

– the integration of statistical techniques for disambiguation with related techniques for learning new grammar rules in the face of parse failure, allowing robust coverage of linguistic phenomena that the grammar cannot currently cover; and

– tackling the problem of evaluating the accuracy of parsing systems, by producing "gold standard" data and automatic evaluation software.

-- SPARKLE: Shallow PARsing and Knowledge extraction for Language Engineering

The first goal of the SPARKLE project was to produce generic software able to reliably produce a unique, correct but simple phrasal-level syntactic analysis of naturally occurring free text, in four European languages. The software was capable of practical use for processing of substantial quantities of text. The second goal was to develop a lexical acquisition system capable of learning from free text certain types of grammatical information about words, such as subcategorisation (e.g how verbs can link up with particular types of phrases), argument structure (how these types of linkages can represent meanings), and semantic selection preferences (biases in what types of meanings go together).

The aim of the project was to deploy and test parsers in multilingual information retrieval and speech dialogue systems. At Sussex we specifically worked on:

– syntactic annotation schemes for corpora and evaluation standards for parsers;

– developing a robust and accurate phrasal parser of English;

– a system for automatically acquiring subcategorisation information from corpora; and

– techniques for modeling semantic type, acquisition of selectional preferences, and automatic recognition of diathesis alternations.

1.4. Representation of Language Data

-- LEXSYS: Analysis of Naturally-occurring English Text with Stochastic Lexicalized Grammars

In the LEXSYS project we hand-crafted a wide-coverage, lexicalized tree grammar, in which each word is associated with one or more `elementary' tree structures (which are combined to produce complete syntactic structures); we also implemented an associated parser that assigns rich descriptions to the sentences it parses, and created a system for structural disambiguation with such grammars. We devised and implemented a number of novel techniques which address a number of important problems in developing large grammars and processing with them. These can be divided into three areas:

Grammar size: we developed techniques for encoding what is logically a single grammar in a variety of different ways, each encoding tailored to a particular task. In doing this, we exploited the fact that the grammar is inherently redundant along certain dimensions in order to substantially reduce problems stemming from its size.

Efficiency: we designed the grammar in a way that addresses the computational problems that typically arise when large structures are used extensively in hand-crafted grammars. The key to this involved localizing those dependencies within the elementary structures of the grammar that a parser is required to check.

Disambiguation: we devised and experimented with a probabilistic technique for acquiring knowledge of which words are able to

function as dependents of others, using a semantic hierarchy to group together senses of nouns into semantically similar classes.

-- Annotation of Language Data: SUSANNE, CHRISTINE and LUCY

The SUSANNE (Surface and Underlying Structural Analysis of Natural English) project designed an annotation scheme for English encoding the detailed phrasal structure of sentences, and produced a 130,000-word corpus of written English annotated in accordance with the scheme. The SUSANNE Corpus is freely available for use by researchers, and has proved to be a very popular and useful resource.

Extending the work of the SUSANNE project, the CHRISTINE Corpus comprises a socially representative annotated sample of current spontaneous speech, applying the annotation standards devised in SUSANNE to create resources for studying structure in present-day British language. It includes various extensions of the annotation scheme to identify the many structural features particular to speech. This corpus is also freely available.

The LUCY project developed an electronic database of structurally analysed modern written English, including not only the "polished" writing of published books and magazines but also the writing of young children and teenagers.

These annotated corpora have been used for many different purposes, including comparing the complexity of adult and child language, analysing differences between spoken and written English, statistical training of automatic language analysis systems, and evaluating the accuracy of such systems.

-- Multilingual Lexicons: PolyLex

Computational linguists have made significant advances over the last dozen years in developing theoretically motivated techniques for representing the lexicons of individual languages. By contrast, little progress has yet been made in the design of lexicons for two or more related languages. However, such multilingual lexicons will be central to the operation of many of the products of the natural language processing industry that will appear in the next two decades.

In the PolyLex project, we developed a trilingual computer lexicon for the core vocabulary of Dutch, English and German. From a linguistic perspective, this allowed us to ascertain the extent to which these Germanic languages can be lexically related, examining formal ways of expressing linguistic generalizations that hold across two or more languages, and assessing the degree to which the historical links between languages can be exploited in descriptions of the languages as they are now. From a computational perspective, we evaluated how well existing techniques for representing monolingual lexicons generalize to the multilingual case and investigated the extent to which multi-language lexical representation techniques may be applicable within monolingual lexicons.

The project developed an inheritance-based trilingual lexicon for the core vocabulary of Dutch, English and German using inheritance networks to share information across the languages at all levels of linguistic description.

2. Language Processing Applications

-- Sentiment Analysis

An automatic system that could accurately determine whether a document expresses positive or negative opinions (also called sentiment analysis) would be useful for a number of different types of user: for instance in brand and corporate image monitoring, investment analysis, product marketing, and consumer research into goods and services. Most approaches to this problem use machine learning techniques, inducing sets of features that usually indicate positive or negative

language, based on hand-classified reviews or news articles.

We are currently investigating a number of different aspects of this task, including:

– the influence of domain and text type on sentiment classification accuracy;

– fine-grained classifications of words and phrases using the linguistic theory of Appraisal;

– integrating retrieval of opinionated texts with classification; and

– cross-lingual sentiment analysis (e.g. reporting in English about opinions originally written in Chinese)

-- Natural Habitats

In the near future, many everyday devices (in the home, the office, etc.) will contain substantial amounts of computing power and will collectively provide a wide variety of networked services. The value of such services will be greatly enhanced if the user is able to `compose' them: link them up in ways that are tailored to their own particular environment. This project is investigating how NLP techniques can help make service composition a possibility for non-technical users, focusing on the development of an interactive service composition tool that uses a natural language interface.

As the trend towards ubiquitous computing technology gathers pace, and the potential benefits of the technology begin to emerge, there is a growing need to make configuration of pervasive environments accessible to non-technical users. If we are all to maintain a sense of being in control of the technology around us, we need to be able to tailor the behaviour of our environments in a straightforward way to make it suit our particular personal needs. This amounts to composing virtual services from the set of actual services that populate our environment. The ability to effectively configure a pervasive environment and compose virtual services is not just a matter of adding useful functionality: without such a capability, many people will find the technology so intrusive that they will not want anything to do with it.

Natural Habitat is an interdisciplinary research project that brings together researchers in the areas of distributed systems, natural language processing and human computer interaction, with the aim of exploring the extent to which an approach centred around the use of natural language processing technology can produce tools that support non-technical users in the task of configuring their pervasive environments. We have completed the development of a running prototype system, providing services such as printers, email and alerts in the context of a virtual notice board.

-- PSET: Practical Simplification of English Text

A number of studies have concluded that improving access by disadvantaged groups of people to written language on the World Wide Web should be a priority. One barrier to accessing written material on the Web is that most of the written material is in English, often employing an extensive vocabulary and a sophisticated style that may make the text difficult or impossible to understand for people for whom English is a foreign language in which they are not fluent, or for people who have language disabilities.

In the PSET project we aimed to help widen access to the Web by building a computer system which takes in English (newspaper) text published on the Web, and outputs a simplified version with broadly similar meaning with, for example, uncommon or unusual words replaced with more common or familiar synonyms, and difficult to follow syntactic constructs replaced with simpler ones (e.g. changing passive constructions to active).

Over the course of the project we gained a good understanding of what types of simplifications could be useful for people with aphasia (a medical condition in which understanding or production of language is impaired, perhaps as a result of a major accident or a stroke), and experimented with a prototype system that could perform these types of simplifications. Unfortunately, we were not able to deploy the system within the constraints of the project, but we gained valuable experience in building the tools necessary to do this.

-- DEEP THOUGHT: Hybrid Deep and Shallow Methods for Knowledge-Intensive Information Extraction

This project was concerned with devising methods for combining robust shallow methods for language analysis with deep semantic processing. Shallow analysis can always find some sort of analysis for a piece of text, but the analysis might not contain enough detail for advanced types of language-based application.

The main idea behind the project was to preserve the advantages of shallow processing while adding more accuracy and depth in a controlled fashion at places where the application has a real demand for such increase in semantic analysis. The goal was the detection of relevant types of information, not full text understanding. Shallow processing enriches a text with annotations (parts of speech, phrases, named entities, simple relations). Deep processing is only called at places where shallow analysis hypothesises relevant relations but cannot detect or select the correct relations. This approach has important advantages. Robustness is maintained. The necessary coverage can be provided by adding to the full-coverage shallow grammars a specialised deep grammar for the relevant domains and semantic relations. Efficiency is guaranteed by adding deep processing to the fast shallow analysis only at places where it is needed and where it has a reasonable chance of producing useful information.

Three knowledge-intensive language-based applications were implemented that were able to benefit from the increased depth in semantic analysis:

– Information extraction for business intelligence;

– Email response management for customer relationship management; and

– Creativity support for document production and collective brainstorming.