Institution: Lancaster University

Unit of Assessment: UoA 11: Computer Science and Informatics

Title of case study: Research on natural language processing leading to improved language tests and dictionaries for millions of language learners

1. Summary of the impact

Worldwide impact on language learners and others has been generated by the development at Lancaster of a ground-breaking natural language processing tool (CLAWS4), and an associated unique collection of natural language data (the British National Corpus, or BNC). Some highlights selected from the primary impacts are as follows:

- Cambridge University Press has significantly improved the quality of its language learning materials (over 60 books) via key enhancements to its core source material that have been enabled by CLAWS4 and the BNC;
- The Society for Testing English Proficiency (STEP) in Japan has based a widely-used Language Test (EIKEN) on data analysed by CLAWS4. The EIKEN test is taken by two million schoolchildren every year.

The pathways to impact have been primarily via consultancy and via licencing of software IP. The impact itself is largely on the language learners—i.e. users of products such as the above. There is a secondary economic impact on a UK SME which has licenced our software.

2. Underpinning research

Context and research problem. Within computer science the research sits under the general heading of *natural language processing* (NLP). Our approach is to collect and analyse machine-readable corpora (large bodies of text), and then to annotate these corpora—at all levels, including morphological, grammatical, syntactic and semantic—using a combination of manual analysis and automatic tagging [4]. This process then enables bootstrapping of more capable tools, more advanced text retrieval, and better statistical analysis, all of which can be exploited in a range of areas as discussed below.

Description of the underpinning research and findings. The enabling research arose from a number of projects and collaborations involving Lancaster's *University Centre for Computer Corpus Research on Language* (UCREL): a centre for computational and corpus research that spans the School of Computing and Communications and the Department of Linguistics and English Language. UCREL has been at the forefront of corpus-based natural language processing research since the early 1980s. Its two contributions of specific relevance to the present case study are as follows:

1. Collection of the British National Corpus (BNC). The BNC, which consists of 100 million annotated words of British written and spoken English, was initially developed by Roger Garside (Senior Lecturer in Computing, retired in 2008) and Professor Geoffrey Leech (now emeritus) who led a team of 15-20 researchers including Paul Rayson (now a Senior Lecturer at Lancaster). The work was completed in 1994 and the BNC is now probably the best-known and most widely-used English corpus in the world.

2. Development of the CLAWS4 part-of-speech tagger. CLAWS, initially conceived in the 1980s, is a tool that performs automatic labelling of words in running text with word class categories (e.g. noun, verb, adjective, adverb). In 1993/94, the tool was significantly enhanced and retrained as "CLAWS4" [2,3,5], using a two-million-word manually-checked sub-corpus, resulting in a much higher degree of accuracy than was previously possible. The key computational research insight in developing CLAWS4 was to apply a combination of statistical techniques and rule-based methods—both informed by careful manual linguistic annotation—to achieve more robust and accurate tagging performance. CLAWS4 is one of the first taggers to use a hybrid approach, resulting in significantly-improved accuracy over earlier approaches (see below).

Research outcomes. As well as the BNC itself, a new frequency dictionary [1] was produced to replace previous out-dated versions from the 1930s and 1940s, and this has now become the standard reference in the area. In addition, the development of CLAWS4 has resulted in a tool that

is robust across a number of different types of written text and speech, with an accuracy rate of 98.5% (as opposed to 90% for previous generation 'dumb' taggers which pick the most common tag; and 77% for earlier rule-based approaches). CLAWS4 has become the de facto standard for part-of-speech tagging of modern English reference corpora.

3. References to the research

Key references:
[1] Leech, G., Rayson, P. and Wilson, A. (2001). Word frequencies in written and spoken English: based on the British National Corpus. London: Longman.
588 Google Scholar citations as of 30 August 2013
[2] Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, pp. 622-628.
COLING is an international conference that uses blind peer reviewing (h5-index 35). 134 citations
[3] Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.
The book went through Longman's peer review and commissioning process, and each chapter was peer reviewed before being accepted. 222 citations.

Other references:
[4] Leech, G. (1993). Corpus annotation schemes. Literary and linguistic computing, 8(4), 275-281.
[5] Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (eds) Using corpora for language research: Studies in the Honour of Geoffrey Leech. Longman, London, pp. 167-180.

4. Details of the impact

Pathways to impact. There were two main pathways to impact:

- Consultancy to industrial partners. Cambridge University Press used the BNC in its English testing material, and used CLAWS4 to analyse and enhance the Cambridge English Corpus. Oxford University Press used the BNC in a top-selling dictionary product.
- Software IP licencing. Software was licensed to two companies: The Society for Testing English Proficiency (STEP) in Japan; and Lexical Learning Ltd (L2), a UK-based SME.

Impact. The impact has been very wide and highly significant. The main area of impact is on language learners: a huge number of learners worldwide have benefitted from a range of Lancaster-enabled or Lancaster-enhanced products. A secondary area is economic impact on the above-mentioned SME. We expand on these two areas in the following:

1. Impact on language learners

(i) Cambridge University Press used the BNC to underpin its "*English for Speakers of Other Languages*" (ESOL) materials: specifically, the *Preliminary English Test Vocabulary List*, and the materials in *Business English* and *Key English*. In 2010, there were nearly 3,500,000 entrants for Cambridge English exams, with more than 11,500 universities, employers and government departments worldwide recognising and using Cambridge English qualifications.

(ii) Cambridge University Press also used CLAWS4 to analyse and enhance its one-billion-word *Cambridge English Corpus*, thus providing enhanced word frequency information and language examples for over 60 dictionaries, course books and other items (e.g. exam materials) related to English Language Teaching.

(iii) The *Oxford Advanced Learner's Dictionary* (OALD) is also based on the BNC. It is the world's best-selling advanced learner's dictionary, now on its 8th edition, and has sold over 35,000,000 copies worldwide (2,000,000 printed copies since 2010).

(iv) The award-winning Japanese TV programme *100-go de START Eikaiwa* ("Let's start English with 100 basic core words") used the BNC as its key source data. This programme, broadcast by NHK, regularly had more than 1,000,000 viewers. The textbooks and CD materials produced

alongside the TV programme sold over 800,000 copies.

(v) The Society for Testing English Proficiency (STEP) in Japan used CLAWS4 to analyse its natural language data in order to assist its teams in preparing better language tests. STEP is Japan's largest English language testing body and delivers its EIKEN tests to more than 2,000,000 school students every year.

(vi) Lancaster's highly-cited word frequency dictionary [1] led directly to an on-going series of word frequency dictionaries for other world languages, co-edited by Rayson and published by Routledge. Within the REF period, the following dictionaries have been published: Chinese (2009), French (2009), American English (2010), Czech (2010), Arabic (2011), Japanese (2013), Russian (2013) and Dutch (2013). These have had significant impact on language teaching and learning communities, with total sales of 17,238 copies as of September 2013.

Reach: Worldwide and huge (see numbers above). Significance: The materials described in (i), (iii), (iv) and (vi) are all directly reliant on the BNC. In (ii), CLAWS4 enables Cambridge editors and authors to refine search results and collocation lists by part-of-speech, significantly improving the quality of the resulting dictionaries, course books, etc., while reducing effort. In (v), CLAWS4 enables STEP's test developers to perform finer-grained searching and thus significantly improve language tests and vocabulary lists.

2. Impact on a UK SME

Lexical Learning Ltd is a UK SME that provides software products for language learning. It has incorporated CLAWS4 into its WORDREADY product to automatically extract part-of-speech information from its source texts. Reach: Since it became available in 2011, WORDREADY has generated subscription sales of more than £10,000. Significance: CLAWS4 gives WORDREADY a commercial edge by enabling the presentation of syntactically-richer questions to learners, thus providing a significantly-enhanced learner experience.

5. Sources to corroborate the impact

1. The EPSRC/BCS/IEE International Computer Science Review in 2001 recognised that UCREL has led the way in an approach to statistical natural language processing based upon information from large bodies of naturally-occurring text.
2. The Global Corpus Manager of Cambridge Dictionaries, Cambridge University Press, has vouched for the impact that the part-of-speech annotation has had on the Cambridge English Corpus and the Cambridge International Dictionary of English and other learner publications.
3. Cambridge University Press website contains details of the books produced using their corpus data http://www.cambridge.org/elt/corpus/
4. Cambridge English Language Tests website with details of the number of students: http://www.cambridgeesol.org/about/news/annual-review-2010.html
5. Oxford website showing the importance of the BNC to dictionary making at OUP http://oald8.oxfordlearnersdictionaries.com/bnc.html
6. The Chief Researcher from the Research & Test Development Section of STEP-EIKEN in Japan is involved in the deployment of CLAWS4 on data used in the language tests.
7. STEP-EIKEN's website contains details of the number of students that take their tests http://stepeiken.org/
8. The Director of Lexical Learning Ltd, described CLAWS4 as one of the key technologies in their WORDREADY Academic English vocabulary tutoring systems which has won a British Council ELTons Award in 2012 in the Digital Innovation category.