Do HCI and NLP Interact?

Karamanis, N.; Schneider, A.; van der Sluis, Ielka; Schlogl, S.; Doherty, G.; Luz, S.

Published in:
Proceedings of the 27th CHI-2009

Publication date:
2009

Do HCI and NLP Interact?

Nikiforos Karamanis
Nikiforos.Karamanis@cs.tcd.ie

Anne Schneider
schneia@cs.tcd.ie

Ielka van der Sluis
ielka.vandersluis@cs.tcd.ie

Stephan Schlogl
schlogls@cs.tcd.ie

Gavin Doherty
Gavin.Doherty@cs.tcd.ie

Saturnino Luz
luzs@cs.tcd.ie

Department of Computer Science
Trinity College Dublin
Dublin 2
Ireland

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-246-7/09/04.

Abstract
We examine the relationship between HCI and Natural Language Processing (NLP) by performing a bibliometric analysis and looking at the specific example of BioNLP. We identify opportunities for HCI to fertilise current NLP research and suggest that HCI will benefit from looking at advances in NLP more closely.

Keywords
Bibliometric analysis, literature review, BioNLP.

ACM Classification Keywords
I.2.7 Natural Language Processing. H.5.0. Information interfaces and presentation (e.g., HCI): General.

Introduction
In their leading textbook on speech and language processing, Jurafsky and Martin (2008) [9] introduce NLP as the field which aims to provide computers with the ability to process human language. The ultimate goal of NLP is to get computers to perform useful language-related tasks such as conversing with a human, translating a document, answering questions using information from the Web, etc [9:35].

Although the term Human-Computer Interaction, is absent in the index of at least two of the most widely used textbooks in NLP [9,11], other members of the NLP community have investigated the relationship between HCI and NLP in more detail. A few years ago, Ozkan and Paris (2002) [12] argued that NLP and HCI

have similar concerns but observed limited interaction between the two fields. Similar remarks were made by Dybkjaer and Bernsen (2000) [8] and Larsen (2003) [10]. HCI researchers have also called for a synthesis between HCI and Artificial Intelligence, which encompasses NLP (see [18] for an overview).

This survey investigates whether NLP and HCI have come any closer since these remarks were made. This is of interest, both in terms of understanding the impact of HCI on other areas, and of ensuring that the HCI community is responding to the challenges introduced by recent advances in NLP.

Overlap between NLP and HCI

Given the wide range of work which falls within the realm of HCI, trying to identify the subset that has impacted on NLP can be challenging: e.g. Do papers discussing speech-enabled interfaces cross the border between the two disciplines by definition? NLP is also quite diverse, giving rise to similar questions: e.g. Does evaluating an NLP component by collecting human judgments or by measuring performance on a task borrow from HCI methods? Or, conversely, have methods employed in NLP evaluation had an impact on HCI methodology?

As a starting point for our exploration and in order to get a general idea of the level of overlap between the two fields, we performed a bibliometric analysis of research in NLP and HCI, extending a preliminary survey by Reiter [13]. Reiter analysed the citations of papers published in two major NLP journals in 2005 to identify which fields have the most impact on recent NLP research.

We extended Reiter's study by extracting additional citations from articles published in 2007 in five major NLP and five major HCI journals (the ones with the highest impact factor).[1] Then, we computed how many times each journal cites (a) itself, (b) the other four journals in the same category and (c) the five journals in the other category. Table 1 presents the results, normalised by the total number of citations to journals in the ISI database.

The Table accords with Reiter's results, showing very limited influence from HCI on NLP. It also shows that work in NLP has had very little influence on HCI. The small amount of cross-citations is mostly related to work on speech and dialogue processing (from *Speech Communication* to the *International Journal of Human-Computer Studies* and vice versa).

The BioNLP case

To make the investigation more focused, we reviewed recent work in BioNLP, the subarea of NLP which is dedicated to the analysis of text in the biomedical domain [2]. We chose this area for three reasons: First, BioNLP investigates problems of general interest in NLP (such as methods for recognising important terms and extracting information from text documents) and has experienced substantial growth in recent years. Second, by looking at BioNLP we focus on text analysis, unlike previous reviews which were interested in text generation [12] and speech and dialogue processing [8,10]. Third, BioNLP researchers have expressed a clear interest in reaching out to a large community of potential users, namely biomedical experts. This

[1] The data are available online by the ISI Web of Knowledge: http://admin-apps.isiknowledge.com/JCR

concern was made explicit in the BioNLP track of the 2008 *Pacific Symposium for Biocomputing* (PSB), one of the main bioinformatics conferences. The track was dedicated to investigating the utility, usability, portability and reliability of BioNLP systems [5]. One would expect work in HCI to be particularly relevant to these topics.

		self-citations	NLP	HCI
NLP	**CL**	54.55%	5.05%	**0.00%**
	CSL	16.09%	43.53%	**0.00%**
	SC	19.86%	18.38%	0.46%
	LRE	7.46%	65.67%	**0.00%**
	IEEE	26.84%	12.37%	**0.00%**
HCI	**HCI**	25.40%	**0.00%**	7.94%
	UMUAI	20.14%	**0.00%**	9.03%
	IJHCS	12.58%	0.77%	8.28%
	IWC	14.36%	**0.00%**	22.67%
	BIT	32.64%	**0.00%**	30.56%

Table 1. Cross citations between five major NLP and HCI journals (data extracted from 2007 issues). The journals considered are Computational Linguistics (CL), Computer Speech and Language (CSL), Speech Communication (SC), Language Resources and Evaluation (LRE), IEEE Transactions on Audio Speech and Language Processing (IEEE), Human-Computer Interaction (HCI), User Modeling and User-Adapted Interaction (UMUAI), International Journal of Human-Computer Studies (IJHCS), Interacting with Computers (IWC) and Behaviour and Information Technology (BIT).

Concerns in BioNLP

NLP output is typically evaluated quantitatively in terms of precision, recall and their harmonic mean (F-score) against answers annotated by humans on the text [9:489]. A primary concern within the BioNLP community is whether this type of intrinsic evaluation is sufficient [5]. To address this issue, several papers in PSB 2008 compare the results of such evaluations with the results of extrinsic evaluations, mainly timing studies measuring human performance on a certain task (such as database curation or online search for information).[2]

Carporaso et al. [4] observe that high results in intrinsic evaluation do not necessarily improve curation performance due to the need to access external sources of information. In a related study, Wang and Matthews [16] added functionalities to an extant curation interface to utilise several approaches for a certain NLP task (term normalisation). One of these approaches, which produces a list of suggestions, is shown to increase curation speed more than the others although it fares worse than them in terms of its F-score. This is because the other approaches often do not return any suggestions thus forcing curators to perform time-consuming searches for the missing information. However, curation is slowed down considerably when the list of suggestions becomes too long.

Given that NLP software components are bound to be imperfect, Alex et al. [1] investigate whether it is worth exposing curators to their flaws. Similarly to Wang and Matthews, they adjusted an existing curation interface to present the output of several NLP processes and measured curation time under three conditions: (a) 100% correct NLP output (provided by human annotation) (b) real NLP output (which contains errors) and (c) no NLP output at all (control condition). These studies indicate that in some cases real NLP output can speed up curation compared to the control condition

[2] From the 29 papers submitted to the BioNLP track, nine were accepted. Four of those report on advanced BioNLP methods or address software engineering issues and are therefore irrelevant for our purposes.

while flawless NLP analysis often leads to additional gains in efficiency. However, there is a lot of variability in the performance of the curators and the authors acknowledge that additional parameters such as accuracy, quality, coverage and agreement between curators need to be considered before any final conclusions can be drawn.

In two supplementary questionnaire studies, Alex et al. observe a preference by curators for recall over precision (which suggests that NLP can be optimised towards one direction at the expense of the other) and for consistent (yet often incorrect) NLP output even though the latter is shown to slow down curation. So there seem to be complex interdependencies between NLP performance in benchmark evaluations with respect to at least the demands of the actual curation task, user preferences and the results of timing studies. This makes it hard to determine under which circumstances NLP is actually useful.

The timing study in Alex et al. points to the second considerable concern of the BioNLP community: How helpful would NLP be if it were 100% correct? To answer this question, Divoli et al. [6] performed a web-based questionnaire study, which indicated that NLP-aided term expansion can be helpful for searching biomedical information online. In a follow-up study, they deployed mock-up prototypes (a sequence of screenshots adjusted from an extant search engine) to confirm the above finding, and to test different ways of implementing term expansion (hyperlinks versus checkboxes) in the interface. Finally, Roberts and Hayes [14] analysed a large number of questions posed to librarians and found out that about 27% of those could be processed using current NLP techniques.

Notably, in the reviewed papers we found only one HCI-related reference[3], a citation to the Shneiderman and Plaisant textbook [16] in Divoli et al. accompanied by an overview of the iterative approach to system development. All other references were papers in BioNLP, NLP, bioinformatics and information sciences.

Relevance to HCI

It strikes us that one useful and immediate contribution HCI methods could make to the evaluative work reviewed above relates to modelling: both at the level of underlying human factors and at the higher level of task analysis. Performance modelling of basic selection tasks could, for instance, be used to inform the design and contextualise the results of studies such as Wang and Matthews'. As regards task analysis, although none of the reviewed papers presents or cites a detailed analysis for the investigated tasks (database curation and search for information), performing such analysis can ensure that the subtasks chosen for extrinsic evaluation are indeed representative of the work that biomedical experts carry out on a daily basis.

In addition to evaluation, HCI methods are relevant from a system design perspective. Introducing user-centered approaches, for instance [3], would shift the focus from adding functionalities to existing interfaces into placing more emphasis on the overall process and context of work. This could shed light on some of the observed complex interdependencies and help clarify under which circumstances NLP does indeed provide added value. Looking at user's strategies to overcome

[3] Primarily, we were looking for citations to the HCI journals in Table 1 and conferences such as CHI, UIST, INTERACT, HCI International, British HCI, Nord/Oz-CHI, IUI, etc.

errors in more detail may provide additional insight with respect to these issues and help feed evaluation back to overall system design.

Thus, we advocate that HCI can fertilise research in NLP by introducing methods such as task and error analysis as well as contextual inquiry, which can provide a sound basis for the development of NLP-enabled systems. This will allow the investigation of additional issues such as assessing the learnability of such systems (which was mentioned in the call for papers in [5] but has not been addressed yet).

However, our bibliometric survey has indicated that HCI researchers are not that familiar with ongoing NLP research, and thus with the opportunities emerging from this field. This is reflected in the discussion of NLP in the leading HCI textbooks. Sharp et al. [15:113-114] briefly discuss the differences between text and speech-based interaction and provide some general design guidelines for language-based interfaces. Dix et al. [7:138-139] contrast language-based interaction with direct manipulation which is considered to be a more attractive alternative. A similar view is held in the more detailed account of NLP research by Shneiderman and Plaisant [16]. However, they also add that HCI studies focused on discovering and analysing the tasks and situations for which NLP-enabled applications are most beneficial can make their use more widespread [16:332]. We make a similar point by emphasising the need to fertilise BioNLP research with contextual design methods.

HCI textbooks view NLP mostly as contributing towards the development of yet another mode of interaction. However, most NLP analysis takes place in the background (e.g. to mine the literature, identify relevant passages, deal with duplicating or contradictory information, etc) and the way in which the retrieved information will be presented to the user does not necessarily have to be in natural language. Moreover, NLP has now begun to support tasks such as database curation or advanced online search which cannot be performed in large scale otherwise. Thus, a suitable system needs to be designed using HCI techniques to incorporate the results of the NLP analysis and assist users with their tasks. As NLP techniques become more mature, this need is likely to become more pressing.

Members of the speech and dialogue processing community [8,10] were between the first ones to raise the need for more interaction between the two fields. Our review indicates that the limited cross-referencing which has taken place in this area does not extend to other NLP subdomains. One interesting question is whether design principles developed to deal with the inaccuracy of speech-based interaction can be applied to tackle the imperfections of text-related NLP, given the differences pointed out by [15].

More generally, it seems that in situations where the use of NLP provides the only reasonable way to accomplish a task, NLP becomes a research challenge for HCI. While standard interaction design assumes a certain amount of component reliability, the inherent inaccuracy of NLP technology (of which NLP researchers are very much aware) might call for new HCI methods to be developed.

Conclusion

In our ongoing work we aim to build on the opportunities identified in this paper and bring NLP and HCI closer to each other. To investigate their relationship in more detail, we will be looking more closely at work in speech and dialogue processing and subareas of NLP other than BioNLP.

The application of contextual techniques for the development of NLP systems appears to be the obvious starting point for our applied work given our analysis. We also want to investigate whether the inherent inaccuracy of NLP systems can motivate new HCI approaches.

Acknowledgements

This research is funded by Science Foundation Ireland as part of the CNGL project (www.cngl.ie).

References

[1] Alex, B., Grover, C., Haddow, B., Kabadjor, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., and Wang, X. Assisted Curation: Does Text Mining Really Help? *Proc. PSB 2008*, 556-567.

[2] Ananiadou, S., and McNaught, J. Text Mining for Biology And Biomedicine. Artech House, 2006.

[3] Beyer H., and Holtzblatt, K. *Contextual Design: Defining Customer-centered Systems*. Morgan Kaufmann, 1998.

[4] Caporaso, G.J., Deshpande, N., Fink, J.L., Bourne, P.E., Cohen, K.B., and Hunter, L. Intrinsic Evaluation of Text Mining Tools May Not Predict Performance on Realistic Tasks. *Proc. PSB 2008,* 640-651.

[5] Cohen, K.B., Yu, H., Bourne, P.E., and Hirschman, L. Translating Biology: Text Mining Tools That Work. *Proc. PSB 2008*, 551-555.

[6] Divoli, A., Hearst, M.A., and Wooldridge, M.A. Evidence for Showing Gene/Protein Name Suggestions in Bioscience Literature Search Interfaces. *Proc. PSB 2008*, 568-579.

[7] Dix, A., Finlay, J. E., Abowd, G. D., and Beale, R. *Human-Computer Interaction*. Prentice Hall, 2004.

[8] Dybkjaer, L., and Bernsen, N.O. Usability Issues in Spoken Dialogue Systems. *Nat. Lang. Eng 6*, (2000), 243-271.

[9] Jurafsky, D., and Martin, J. *Speech and Language Processing*. Prentice Hall, 2008.

[10] Larsen, L. B. Assessment of Spoken Dialogue System Usability - What Are We Really Measuring? *Proc. EUROSPEECH 2003*, 1945-1948.

[11] Manning, C.D., and Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[12] Ozkan, N., and Paris, C. Cross-fertilization Between Human Computer Interaction and Natural Language Processing: Why and How. *International Journal of Speech Technology 5*, 2 (2002), 135-146.

[13] Reiter, E. The Shrinking Horizons of Computational Linguistics. *Comput. Linguist. 33,* 2 (2007), 283-287.

[14] Roberts, P.M., and Hayes, W.S. Information Needs and the Role of Text Mining in Drug Development. *Proc. PSB 2008*, 592-603.

[15] Sharp, H., Rogers, Y., and Preece, J. Interaction Design: Beyond Human-Computer Interaction. Wiley, 2007.

[16] Shneiderman, B., and Plaisant, C. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley, 2005.

[17] Wang, X., and Matthews, M. Comparing Usability of Matching Techniques for Normalising Biomedical Named Entities. *Proc. PSB 2008*, 628-639.

[18] Winograd, T. Shifting Viewpoints: Artificial Intelligence and Human-Computer Interaction. *Artificial Intelligence 170*, (2006) 1256–1258.