Gokhan Tur, Ye-Yi Wang,
and Dilek Hakkani-Tür

TechWare: Spoken Language Understanding Resources

In this issue, "Best of the Web" focuses on spoken language understanding (SLU), an emerging field in between the areas of speech processing and natural language processing. Since SLU is not a single stand-alone technology, unlike speech recognition or synthesis, it is hard to present a single application. The term *spoken language understanding* has largely been coined for targeted understanding of human speech directed at machines, although understanding human/human conversations or even human/human/machine interactions are vibrant areas of research. For a more comprehensive survey of SLU tasks, readers are referred to [1].

Typically, SLU tasks and the approaches are quite different for each application and environment (such as mobile device versus television). There is also a strong interest from the commercial world in SLU applications. They typically employ knowledge-based approaches (mainly based on handcrafted grammar rules or finite set of commands) and are now used in some environments such as smartphones, cars, call centers, and robots. The state-of-the-art approaches rely on data-driven methods and are heavily used in academic and industrial research labs, though these methods started to propagate to commercial applications like mobile personal assistants.

This column first presents a very high-level review of the SLU technology, starting from its place in a spoken dialog system, then focusing on well-established SLU tasks such as domain detection, intent determination, and slot filling, along with corresponding benchmark data sets and methods.

SLU OVERVIEW

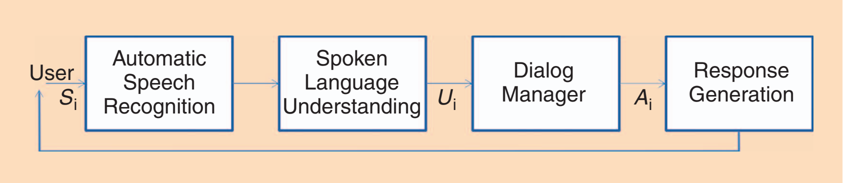
At a very high level, the basic components of a spoken dialog system are shown in Figure 1. The goal of understanding is to convert the recognition of user input, S_i, into a task-specific semantic representation of the user's intention, U_i at each turn. The dialog manager then interprets U_i and decides on the most appropriate system action, A_i, exploiting semantic context, user-specific meta-information, such as geolocation and personal preferences, and other contextual information. For example, if the user clicks/touches on a map on the screen and says "How much is the cheapest gas around here?" the system should be able to interpret the domain, intent, and the associated arguments, such as:

- *Domain*: Local Business;
- *Intent*: Get_Price
- *Slots*: {*good*: gas; *cost_relative*: cheapest; *location*: (lat,long)}

IN THIS COLUMN, WE FOCUS ON THE KEY TASKS OF A SLU SYSTEM AS USED IN HUMAN/MACHINE CONVERSATIONAL SYSTEMS.

In this column, we focus on the key tasks of a SLU system as used in human/machine conversational systems. These include domain detection, intent determination, and slot filling. We will mainly cover data-driven techniques. Typically, word n-grams are used as features after preprocessing with generic entities (e.g., dates, locations, or phone numbers) or domain-specific entities (e.g., airline names or airport locations for the flight domain). For generic-named entity extraction, one can use an already available parser, which supports monocase and lack of punctuation (e.g., the Stanford parser: http://nlp.stanford.edu/software/CRF-NER.shtml) or retrain one using the available corpora released by Linguistic Data Consortium (LDC), such as the Automatic Content Extraction (ACE) corpus (http://projects.ldc.upenn.edu/ace) or the Ontonotes corpus (http://www.bbn.com/ontonotes). For domain-specific named entity extraction, one can exploit the knowledge bases, such as Freebase (http://www.freebase.com). Such Semantic Web resources are also proven to be very effective for SLU since they also provide relations between the entities (e.g., "movie-name is-directed-by movie-director") [2], [3].

For building context-free grammar (CFG)-based or Voice XML-based conversational systems, readers are referred to [4], providing examples using the



[FIG1] A conceptual architecture of a spoken dialog system.

Digital Object Identifier 10.1109/MSP.2013.2241313
Date of publication: 5 April 2013

Oregon Health and Science University (OHSU) Center for Spoken Language Understanding (CSLU) Toolkit (http://www.cslu.ogi.edu/toolkit). Similarly, the CMU Phoenix context-free grammar (CFG) parser [5] can be used for semantic parsing of natural language input (http://wiki.speech.cs.cmu.edu/olympus/index.php/Phoenix).

SLU DATA SETS

We provide descriptions and links for two well-known SLU data sets, Airline Travel Information System (ATIS) [6], [7] and MEDIA [8]. Besides these, readers may want to check the HCRC Map Task (LDC Catalog LDC93S12) and University of Rochester TRAINS (LDC Catalog LDC95S25) corpora for goal-oriented human/human conversations, and the Communicator corpus (LDC Catalog LDC2004T16) for dialog modeling. A more recent data set for SLU from the University of Cambridge is available at http://mi.eng.cam.ac.uk/research/dialogue/corpora for the tourist information domain [9].

ATIS

The most well-known SLU benchmark data set is the Defense Advanced Research Program Agency (DARPA) sponsored ATIS corpus (LDC Catalog LDC95S26). The ATIS task consisted of spoken queries on flight-related information. An example utterance is "I want to fly to Boston from New York next week." Understanding was then reduced to the problem of extracting the intent, such as "Flight Info" or "Ground Transportation" and task-specific arguments, such as "Destination" and "Departure Date." The training set contains 4,978 utterances selected from the Class A (context independent) training data in the ATIS-2 and ATIS-3 corpora, while the test set contains 893 utterances from the ATIS-3 Nov93 and Dec94 data sets. Each utterance has its named entities marked via table lookup, including domain specific entities such as city, airline, airport names, and dates. In total, there are 17 intents (goals) and 78 slots.

MEDIA

The EVALDA/MEDIA project originally started in France, focusing on tourist and hotel information. They have adopted a hierarchical semantic structure, providing a more expressive representation than ATIS and allowing the sharing of substructures. For example, the "Flight" subframe can be shared by both "Flight Info" and "Flight Status" frames. On the other hand, the flat concept representation is simpler and often results in a simpler model. The MEDIA data is available via the European Language Resource Association (ELRA): ELRA Catalog ELRA-S0272. The more recent multilingual LUNA SLU project sponsored by the European Union has adopted a different representation inspired from the ICSI FrameNet (https://framenet.icsi.berkeley.edu/) project for better generalization, and the MEDIA corpus has been reannotated accordingly [10].

SEMANTIC UTTERANCE CLASSIFICATION

The semantic utterance classification tasks of domain detection and intent determination aim at classifying a given speech recognition output, X_i, into one of M semantic classes, $\hat{C}_i \in \mathcal{C} = \{C_1, \ldots, C_M\}$. Formally,

$$\hat{C}_i = \arg\max_{C_i} P(C_i | X_i). \qquad (1)$$

While the traditional solution to semantic classification is the bag-of-words approach as used in information retrieval, with the advances in machine learning, researchers have, in the last decade, started employing discriminative classification techniques for this task.

Because of the very large dimensions of the input space, large-margin classifiers like support vector machines [e.g., SVMLight (http://svmlight.joachims.org)], or Boosting [e.g., ICSIBoost (http://code.google.com/p/icsiboost)] were found to be very good candidates. For evaluation, either the top class error rate or F-measure metrics are reported and recall/precision or receiver operating characteristic (ROC) curves are drawn.

SLOT FILLING

The state-of-the-art method for slot filling is framing it as a sequence tagging problem (even for hierarchical semantic representations) and employing corresponding statistical techniques such as hidden Markov model (HMM) or more recently conditional random fields (CRFs) [e.g., Wapiti (http://wapiti.limsi.fr) and CRF++ (http://crfpp.googlecode.com/svn/trunk/doc)].

More formally, the most probable slot sequence is obtained as

$$\hat{Y} = \operatorname*{argmax}_{Y_i} p(Y_i | X_i),$$

where $X_i = x_1, \ldots, x_T$ is the word sequence and $Y = y_1, \ldots, y_T$, $y_i \in C$ is the sequence of associated class labels, C.

CRFs are shown to outperform other classification methods for sequence classification [11], since the training can be done discriminatively over a sequence with utterance level optimization. Similar to maximum entropy models, in this model, the conditional probability, $p(Y|X)$ is defined as

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_k \lambda_k f_k(y_{t-1}, y_t, x_t)\right)$$

with the difference that both X and Y are sequences instead of individual local decision points given a set of features f_k (such as n-gram lexical features, state transition features, or others) with associated weights λ_k. $Z(X)$ is the normalization term. After the transition and emission probabilities are optimized, the most probable state sequence, $\hat{Y}$, can be determined using the well-known Viterbi algorithm.

Usually, in/out/begin representation is employed, following the named entity extraction literature, and slot level F-measure is computed (e.g., http://www.cnts.ua.ac.be/conll2000/chunking/output.html).

AUTHORS

Gokhan Tur (gokhan.tur@ieee.org) is a senior researcher at Conversational Systems Research Center, Microsoft Research, Mountain View, California.

Ye-Yi Wang (yeyiwang@microsoft.com) is a principal research manager at Microsoft Online Services Division, Bellevue, Washington.

Dilek Hakkani-Tür (dilek@ieee.org) is a senior researcher at Conversational Systems Research Center, Microsoft Research, Mountain View, California.

REFERENCES

[1] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Chichester, U.K.: Wiley, 2011.

[2] G. Tur, M. Jeong, Y.-Y. Wang, D. Hakkani-Tür, and L. Heck, "Exploiting the Semantic Web for unsupervised natural language semantic parsing," in *Proc. Int. Conf. Spoken Language Processing (Interspeech)*, Portland, OR, Sept. 2012.

[3] L. Heck and D. Hakkani-Tür, "Exploiting the Semantic Web for unsupervised spoken language understanding," in *Proc. IEEE Spoken Language Technologies (SLT) Workshop*, Miami, FL, Dec. 2012.

[4] M. F. McTear, *Spoken Dialogue Technology: Towards the Conversational User Interface*. London: Springer, 2004.

[5] W. Ward and S. Issar, "Recent improvements in the CMU spoken language understanding system," in *Proc. ARPA Human Language Technology Conf. (HLT)*, Mar. 1994, pp. 213–216.

[6] P. J. Price, "Evaluation of spoken language systems: The ATIS domain," in *Proc. DARPA Workshop Speech and Natural Language*, Hidden Valley, PA, June 1990.

[7] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in ATIS?," in *Proc. IEEE Spoken Language Technologies (SLT) Workshop*, Berkeley, CA, 2010.

[8] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the French MEDIA dialog corpus," in *Proc. Int. Conf. Spoken Language Processing (Interspeech)*, Lisbon, Portugal, Sept. 2005.

[9] M. Henderson, M. Gasic, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young, "Discriminative spoken language understanding using word confusion networks," in *Proc. IEEE Spoken Language Technologies (SLT) Workshop*, Miami, FL, Dec. 2012.

[10] S. Hahn, M. Dinarelli, C. Raymond, F. Lefevre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing stochastic approaches to spoken language understanding in multiple languages," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 19, no. 6, pp. 1569–1583, 2011.

[11] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proc. Int. Conf. Spoken Language Processing (Interspeech)*, Antwerp, Belgium, 2007.

[applications CORNER] continued from page 182

achieves higher apparent resolution than pixel-based scheme leading to sharper images, due to relatively smaller overlapping of center spectrum and neighboring aliasing spectra. In particular, the subpixel-based approach is well suitable for font rendering to achieve better reconstruction of sloping edges of the fonts than pixel-based approach (suffering staircase artifacts), as shown in Figure 4(d) and (e).

REMAINING CHALLENGES AND FUTURE WORK

Subpixel rendering and subpixel-based subsampling are effective in achieving higher apparent luminance resolution than pixel-based schemes with the cost of chrominance distortion. Although a large number of methods have been proposed to deal with the color fringing artifacts, the simultaneous preservation of extra apparent luminance is still an open challenge.

Moreover, most of aforementioned algorithms are designed for conventional horizontal subpixel-based subsampling. Researchers typically do not attempt to apply subpixel-based subsampling to vertical subsampling, as there is a common conception that little can be gained in the vertical direction due to the horizontal arrangement of the subpixels. Consequently, the development of LCD/OLED with "multiprimary" subpixels such as Pentile-RGBW display, VP-RGBW display, and SHARP-RGBY Quattron display has intensified to achieve subpixel resolution both horizontally and vertically. Exploiting efficient and effective encoding/decoding schemes to adapt these "multiprimary" subpixel components is a promising direction for future work.

AUTHORS

Lu Fang (fanglu@ustc.edu.cn) is an associate professor with the University of Science and Technology of China.

Oscar C. Au (eeau@ust.hk) is a professor in the Department of Electronic and Computer Engineering and is the director of the Multimedia Technology Research Center at the Hong Kong University of Science and Technology.

Ngai-Man Cheung (ngaiman_cheung@sutd.edu.sg) is currently an assistant professor with Singapore University of Technology and Design. His research interests are image, video, and signal processing.

REFERENCES

[1] L. Fang, O. C. Au, K. Tang, and X. Wen, "Increasing image resolution on portable displays by subpixel rendering: A systematic overview," *APSIPA Trans. Signal Inform. Processing*, vol. 1, no. 1, pp. 1–10, Jan. 2013.

[2] S. Gibson. (2006, July 21). The origins of subpixel font rendering [Online]. Available: http://www.grc.com/ctwho.htm

[3] S. Gibson. (2010, May 28). Sub-pixel font rendering technology [Online]. Available: www.grc.com/cleartype.htm

[4] C. H. B. Elliot, "Active matrix display layout optimization for subpixel image rendering," in *Proc. 1st Int. Display Manufacturing Conf.*, Sept. 2000, pp. 185–187.

[5] R. Lai. (2012, May 3). Under the microscope: Samsung Galaxy SIII's HD super AMOLED display [Online]. Available: http://www.engadget.com/2012/05/03/galaxy-s-iii-microscope-hd-super-amoled/

[6] C. Betrisey, J. F. Blinn, B. Dresevic, B. Hill, G. Hitchcock, B. Keely, D. P. Mitchell, J C. Platt, and T. Whitted, "Displaced filtering for patterned displays," in *Proc. SID Int. Symp. Dig. Tech. Papers*, 2000, vol. 31, pp. 296–301.

[7] Microsoft Corporation. (2002, Jan. 16). ClearType information [Online]. Available: www.microsoft.com/typography/cleartypeinfo.mspx

[8] S. J. Daly, "Methods and systems for improving display resolution in images using sub-pixel sampling and visual error filtering," U.S. Patent 6775420, Dec. 2000.

[9] L. Fang, O. C. Au, K. Tang, X. Wen, and H. Wang, "Novel 2-D MMSE subpixel-based image downsampling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 740–753, May 2012.

[10] L. Fang, O. C. Au, K. Tang, and A. K. Katsaggelos, "Anti-aliasing filter design for subpixel downsampling via frequency domain analysis," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1391–1405, Mar. 2012.

[11] R. C. Gonzalez and E. R. Woods, *Digital Image Processing*. Beijing, China: Publishing House of Electronics Industry, 2005, pp. 420–450.

[12] J. S. Kim and C. S. Kim, "A filter design algorithm for subpixel rendering on matrix displays," in *Proc. 15th European Signal Processing Conf. (EUSIPCO)*, Sept. 2007, pp. 1487–1491.