Socio-Cognitive Language Processing for Special User Groups

Björn W. Schuller and Michael F. McTear

5.1 Introduction

Sociocognitive Language Processing (SCLP)[1] can be considered as a term that covers 'soft factors' in communication. Likewise, it can be seen as a specific kind of and arguably also as an extension to the more traditional and broadly defined field of Natural Language Processing (NLP)—the idea of coping with everyday language, including slang and multi-lingual phrases and cultural aspects, and in particular, with irony/sarcasm/humour, as well as paralinguistic information such as the physical and mental state and traits of the dialogue partner (e.g., affect, age groups, personality dimensions), and social aspects. Additionally, multimodal aspects such as facial expression, gestures or bodily behaviour should ideally be included in the analysis wherever possible. At the same time, SCLP can render future dialogue systems more 'chatty' by not only appearing natural but also by being truly emotionally and socially competent, ideally leading to a more symmetrical dialogue. To do this, the computer should itself have a 'need for humour', an 'increase of familiarity', etc., i.e., enabling computers to experience or at least better understand emotions and personality so that that they have 'a feel' for these concepts. Beyond these ideas, the broader idea of SCLP includes verbal behaviour analysis, a closer coupling between language understanding and generation incorporating social and affective information, and new language resources to meet these ends. In this way, SCLP links NLP expertise more closely with that of psychology, the social sciences, and related disciplines.

[1] As the field of *Spoken* Language Processing is usually abbreviated as SLP, we suggest S*C*LP as a short notation for Sociocognitive Language Processing.

B. W. Schuller (✉)
GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK
e-mail: bjoern.schuller@imperial.ac.uk

M. F. McTear
Computer Science Research Institute, Ulster University, Northern Ireland, UK
e-mail: mf.mctear@ulster.ac.uk

In this short paper, we will exemplify the principle by focusing on the analysis side of NLP, known as Natural Language Understanding (NLU). Further, we will limit the example to *spoken* language understanding (SLU). The principles do, however, similarly apply to Natural Language Generation (NLG) in any form.

5.2 Spoken Language Understanding

In SLU, the text output by a speech recognition system is analyzed in order to determine its meaning. This meaning representation can then be used in a spoken dialogue application or in other tasks such as speech mining, speech information retrieval, or speech translation.

Three stages can be distinguished in the development of computational approaches to language understanding. Up until the late 1980s, developers handcrafted grammars that consisted of rules covering all predicted inputs along with a parser that applied the grammar rules to the input to determine its constituent structure. In this approach, the focus was on the analysis of written texts i.e. natural language understanding (NLU). In the late 1980s, a paradigm shift occurred in which probabilistic and data-driven models that had already been deployed successfully in speech recognition were now applied to language understanding. In this approach, attention turned to spoken language understanding (SLU), as the statistical methods were more able to deal with the ill-formed input typical of spontaneous speech. By around 2006, a new approach was emerging with the application of deep learning neural models to language understanding and the use of end-to-end architectures that eliminated the need for the traditional components of pipelined architectures.

5.2.1 Rule-Based Approaches

In theoretical computational linguistics, language understanding involved two stages of analysis:

1. *Syntactic analysis*—to determine the constituent structure of the input.
2. *Semantic analysis*—to determine the meanings of the constituents.

This two-step approach is based on the *principle of compositionality*, which states that the meaning of a complex expression is determined by the meanings of its constituent parts and the rules used to combine them. In this approach, fine-grained distinctions can be captured through a deep analysis of constituent structure that has a direct bearing on the semantic analysis. For example, there is only one word that is different in the following sentences, but changing the words results in different meaning representations:

- S1: *List all employees of the companies who are based in the city centre*
- S2: *List all employees of the companies that are based in the city centre*

The interpretation of S1 asks for a listing of employees who are based in the city centre while the interpretation of S2 asks for a listing of employees who are not necessarily based in the city centre, but who work for companies based there. This difference can only be picked up by an analysis that reflects the difference between the use of *who* and *that* in these sentences.

Following the syntactic analysis, the syntactic constituents are transformed into a meaning representation that typically takes the form of a logic-based formalism. Using logic in this way offers a deeper level of understanding, as it enables the application of standard mechanisms for inference. For a detailed account of the formalisms used in rule-based NLU, see [14].

In an alternative approach, semantic analysis is performed directly on the input using a *semantic grammar*. Although this approach is not theoretically motivated, it has been applied successfully in dialogue systems to analyze the user's inputs. Semantic grammars are more robust to the sorts of ungrammatical input and recognition errors that occur in spontaneous speech as they focus on the keywords in the input and do not have to analyze every word. However, they are usually domain-specific, so that separate grammars are required for each new domain.

5.2.2 Statistical Approaches

With the emergence of spoken dialogue systems in the late 1980s and early 1990s, it became apparent that the rule-based approaches used extensively in NLU were not sufficient for spoken language understanding (SLU). Whereas the input to NLU was well-formed written text, in spoken dialogue systems, SLU was required to analyze spoken text that did not necessarily follow the same grammar rules, as it often contained self-corrections, hesitations, repetitions, and other types of dysfluency. Moreover, the output of a speech recognition system took the form of a stream of tokens with no structure information such as punctuation to help determine sentence boundaries and structure within the sentence. In the rule-based approach, the input has to match the rules exactly and any small variations require additional rules in order to be accepted by the grammar. In this respect, the statistical approach is more robust as it can handle input that is potentially ill-formed as well as synonymous strings that should result in the same meaning representation, but would require additional rules in the rule-based approach.

In statistical SLU, the focus was mainly on analyzing the language produced in domain-specific task-based dialogue systems—for example, flight reservations, hotel bookings, etc. With the increasing availability of large corpora of spoken dialogues, it became possible to use machine learning techniques to automatically learn mappings between spoken inputs and the required meaning representations from labelled training data. Given the nature of the tasks, it was not necessary to derive a logic-based

meaning representation. Instead, the output of SLU in a spoken dialogue system is typically a frame-based representation consisting of sets of attribute-value pairs that capture the information in an utterance that is relevant to the application [17]. Generally, three elements are extracted: the *domain* to which the utterance relates (for example, flight reservations), the user's *intent* (for example, to book a flight), and the *entities* that are required to make the reservation (for example, **destination**, **date of travel**, etc.). Thus, the representation of an utterance such as *book a flight to London on Friday* would be something like: *domain = flight_reservations*; *intent = book_flight*; *entities: destination = London, day = Friday*. These elements are extracted from the input using machine learning methods such as classifiers (e.g., support vector machines) for the domain and intent, and Conditional Random Fields for the entities [38].

5.2.3 Deep Learning Approaches

Since around 2006, deep learning techniques have been applied in NLU and have been shown to outperform other machine learning-based approaches. For example, [17] used Recurrent Neural Networks for the identification of intents and the extraction of entities, while [13] used a bi-directional RNN to jointly classify intents and extract entities.

Deep learning for SLU makes use of the sequence-to-sequence (Seq2Seq) model, in which, the input and the output are represented as a sequence [33, 36]. The model consists of an encoder and a decoder. The encoder processes the input creating a vector known as a context (or thought) vector that represents the final hidden state of the encoder. The decoder takes this vector and uses it to create the output. Seq2Seq has been used in machine translation to perform a transduction from an input in a source language such as English to an output in a target language such as German, and in dialogue systems to perform a transduction from an input utterance to a response. Encoding involves the use of neural networks, usually recurrent neural networks (RNNS), long short-term memory networks (LSTMs), gated recurrent units (GRUs), and more recently transformer networks. Decoding takes one element at a time to produce an output sequence using the context vector where the word that is generated at each time step is conditioned on the word generated by the network at the previous time step, as well as the hidden state, providing the context from the previous time step.

There are many variations on the encoder-decoder architecture. For example, the *attention mechanism* was introduced to address the problem that performance decreases as the input sequence becomes longer [3]. The more recently introduced Transformer architecture uses attention mechanisms to draw global dependencies between the input and the output [40]. Whereas a traditional encoder treats every item in a sequence as equally relevant, the transformer selects which parts to include in the encoding as a basis for the current prediction. For more details on deep learning in SLU, see [37].

5.2.4 *Implications for Socio-Cognitive Language Understanding*

Currently, SLU focuses primarily on analyzing the textual form of a message and on outputting a representation of its propositional content. However, additional information is conveyed in the prosodic features of a spoken utterance—its phrasing, pitch, loudness, tempo, and rhythm—that can indicate differences in the function of an utterance as well as expressing emotional aspects such as anger or surprise [34]. Other information to support the interpretation of an utterance may come from sensors that provide data about the environmental context, biosensors that report on the user's physical and emotional state, and machine vision systems that can detect non-verbal accompaniments of speech, such as gestures and facial expressions.

Recently researchers in neural dialogue have started to explore how to integrate information about emotional aspects into their models. Ghosh et al. [11] present an extension to an LSTM language model for generating conversational text that was trained on conversational speech corpora. The model predicts the next word in the output conditioned not only on the previous words, but also on an affective category that infers the emotional content of the words. In this way, the model is able to generate expressive text at various degrees of emotional strength. Zhou et al. [42] developed a conversational model called *Emotional Chatting Machine (ECM)* that produces emotion-based responses to any user input. See also [1] for discussion of a study, in which affective content was incorporated into LSTM encoder-decoder neural dialogue models enabling them to produce emotionally rich and more interesting and natural responses.

Adding additional multimodal input streams and integrating them, if required, into a single meaning representation presents challenges that have yet to be addressed. Similarly, generating output that makes use of this richer information is beyond the current state of the art. We discuss the challenges for Sociocognitive Language Processing of adopting these additional features in the next section.

5.3 The Sociocognitive View

Above (cf. Sect. 5.2), we introduced SLU and showed that, in principle, works in this field *do* consider prosody, and often also multimodal information such as considering the facial expression of speakers. One may thus ask what makes the term SCLP different or justified. In [12], the authors find in two experiments that it seems plausible to consider language understanding as "a special case of social cognition". This is based on a model to predict an "interaction between the speaker's knowledge state and the listener's interpretation" [12]. Similarly, the authors in [9] attest the high relevance of "social, cognitive, situational, and contextual aspects" when dealing with language. Further, the fields of Affective Computing (AC) [25], Social Signal Processing (SSP) [24, 41] or Behavioural Signal Processing (BSP) [20], sug-

gest consideration of affective or emotional and social cues for computing systems used, e .g., in human–computer interaction [21], or human dialogue analysis. In particular, in speech analysis [2, 7, 28, 30] and synthesis [18, 19], such information was considered rather early. Further, sub-fields of NLP deal with such information, most noteworthy the discipline of Sentiment Analysis (SA) [6, 15, 22, 32, 39]. Including also acoustic speech feature information, Computational Paralinguistics (CP) [31] provides a broader view on speaker states and traits beyond sentiment, emotion, or social signals including also biological traits such as age, gender, height, or race, personality traits, or health-related state and trait information (cf. also [10, 23]) alongside physiological states such as eating or exercising, and cognitive load.

Likewise, AC and SSP/BSP provide a general beyond-language paradigm for computer analysis and synthesis of affective and behavioural cues—each of which focusing on one of emotional or social intelligence; SA (and related sub-disciplines of NLP such as Opinion Mining) focus on the analysis of a single aspect—here sentiment—and neglect the synthesis side, and CP deals mainly with speech and language analysis and synthesis without a linking model or component such as a dialogue model. This makes NLP the definition that comes arguably closest to SCLP; however, while it aims to deal with 'natural' language, its emphasis on the soft factors of communication is rather weak. By SCLP, we advocate a strong link between language and the authoring or speaking person's state and trait as related to the spoken content and conveyed 'message'. Likewise, a decision such as in the example in Sect. 5.2 regarding the interpretation of S1 potentially in the sense of S2 could be supported by estimation of the social and cultural background of the speaker such as 'native speaker' (or not) or the personality such as 'conscientious' (then, likely taking it for the actual sense of S1).

5.4 Conclusion

In this short contribution, we introduced the term of Sociocognitive Language Processing (SCLP). We further motivated its introduction reviewing key relevant literature in a constructive and synthetic manner with the aim to highlight borders between related existing disciplines and terms such as NLP and SA, and at the same time confine SCLP as compared to broader fields such as AC and SSP.

We exemplified the considerations by language analysis looking at NLU and leading to a sociocognitive view. In the same vein, in NLG, the sociocognitive view lends more weight to the 'soft factors' in communication, such as synthesizing irony, or non-verbal fillers, and behaviours that fit these and accompany the linguistic content. We believe that, by integrating SCLP principles, one can render future dialogue systems more 'chatty' making them not only feel 'natural', but truly emotionally and socially competent (cf., e.g., [8, 29]). Ideally, this will lead to a more 'symmetrical' dialogue, where both ends—humans and computer systems or intelligent machines—will integrate and comprehend soft factors in the communication.

Arguably, for that, communicative technical systems should themselves have a 'need for humour' [5, 26, 35], an 'increase of familiarity' during repeated or prolonged interactions, etc. In other words, it appears required for genuine SCLP to enable computers to experience and have or at least better understand emotions and personality, such that they have 'a feel' for these concepts (cf., e.g., [4]). For example, the degree of conscientiousness of a system might be the decisive factor between taking S1 in our example in the sense of S1 or rather S2 in addition to its interpretation of 'what the user is like' (based also on increased familiarity) in relation to the best interpretation. This will, however, need to be further expanded upon in follow-up considerations and studies including 'Sociocognitive Dialogue Processing'. Beyond these ideas, the broader idea of SCLP includes verbal behaviour analysis, a closer coupling between language understanding and generation incorporating social and affective information, and new language resources to meet these ends. By that, SCLP unites expertise from psychology and social sciences with NLP on the way to enable genuine conversational dialogue systems [16] or emotionally and socially aware computer-mediated communication [27].

Acknowledgements The first author acknowledges funding from the ERC under grant agreement no. 338164 (iHEARu), and the European Union's Horizon 2020 Framework Programme under grant agreements nos. 645378 (ARIA-VALUSPA), 644632 (MixedEmotions), and 645094 (SEWA).

References

1. Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L.: Affective neural response generation. In: European Conference on Information Retrieval, pp. 154–166. Springer (2018)
2. Bachorowski, J.A., Owren, M.J.: Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. Psychol. Sci. **6**(4), 219–224 (1995)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014). arXiv:1409.0473
4. Ball, J.E., Breese, J.S.: Modeling and projecting emotion and personality from a computer user interface (2001). US Patent 6,212,502
5. Binsted, K., Bergen, B., O'Mara, D., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., O'Mara, D.: Computational humor. IEEE Intell. Syst. **21**(2), 59–69 (2006)
6. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. Mag. **28**(2), 15–21 (2013)
7. Dellaert, F., Polzin, T., Waibel, A.: Recognizing emotion in speech. In: Proceedings of the Fourth International Conference on Spoken Language Processing, vol. 3, pp. 1970–1973. IEEE (1996)
8. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.P.: Simsensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, pp. 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems (2014)
9. Dietrich, R., Graumann, C.F.: Language Processing in Social Context. Elsevier, Amsterdam (2014)

10. Furnham, A.: Language and personality. In: Giles, H., Robinson, P. (eds.) Handbook of Language and Social Psychology. Wiley, Hoboken (1990)
11. Ghosh, S., Chollet, M., Laksana, E., Morency, L.P., Scherer, S.: Affect-lm: a neural language model for customizable affective text generation (2017). arXiv:1704.06851
12. Goodman, N.D., Stuhlmüller, A.: Knowledge and implicature: modeling language understanding as social cognition. Top. Cogn. Sci. **5**(1), 173–184 (2013)
13. Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y.N., Gao, J., Deng, L., Wang, Y.Y.: Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In: Interspeech, pp. 715–719 (2016)
14. Jurafsky, D., Martin, J.H.: Speech and Language Processing (3rd ed. draft) (2020). https://web.stanford.edu/~jurafsky/slp3/
15. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
16. McTear, M.F.: Spoken dialogue technology: enabling the conversational user interface. ACM Comput. Surv. (CSUR) **34**(1), 90–169 (2002)
17. Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., Zweig, G.: Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans. Audio, Speech, Lang. Process. **23**(3), 530–539 (2015)
18. Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am. **93**(2), 1097–1108 (1993)
19. Murray, I.R., Arnott, J.L.: Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Commun. **16**(4), 369–390 (1995)
20. Narayanan, S., Georgiou, P.G.: Behavioral signal processing: deriving human behavioral informatics from speech and language. Proc. IEEE **101**(5), 1203–1233 (2013)
21. Paiva, A.: Affective Interactions: Toward a New Generation of Computer Interfaces?. Springer, Berlin (2000)
22. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, pp. 79–86. ACL (2002)
23. Pennebaker, J.W., Graybeal, A.: Patterns of natural language use: disclosure, personality, and social integration. Curr. Dir. Psychol. Sci. **10**(3), 90–93 (2001)
24. Pentland, A.S.: Social signal processing [exploratory dsp]. IEEE Signal Process. Mag. **24**(4), 108–111 (2007)
25. Picard, R.W., Picard, R.: Affective Computing, vol. 252. MIT Press, Cambridge (1997)
26. Ritchie, G.: Current directions in computational humour. Artif. Intell. Rev. **16**(2), 119–135 (2001)
27. Riva, G.: The sociocognitive psychology of computer-mediated communication: the present and future of technology-based interactions. Cyberpsychology Behav. **5**(6), 581–598 (2002)
28. Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T.: Vocal cues in emotion encoding and decoding. Motiv. Emot. **15**(2), 123–148 (1991)
29. Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Maat, M.T., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., Wöllmer, M.: Building autonomous sensitive artificial listeners. IEEE Trans. Affect. Comput. **3**(2), 165–183 (2012)
30. Schuller, B.: Towards intuitive speech interaction by the integration of emotional aspects. In: Proceedings IEEE International Conference on Systems, Man and Cybernetics, vol. 6. IEEE, Yasmine Hammamet, Tunisia (2002). 6 pages
31. Schuller, B., Batliner, A.: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley, Hoboken (2013)
32. Schuller, B., Mousa, A.E.D., Vasileios, V.: Sentiment analysis and opinion mining: on optimal parameters and performances. WIREs Data Min. Knowl. Discov. **5**, 255–263 (2015)
33. Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., Bengio, Y.: Towards end-to-end spoken language understanding. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5754–5758. IEEE (2018)

34. Shriberg, E., Stolcke, A.: Prosody modeling for automatic speech recognition and understanding. In: Mathematical Foundations of Speech and Language Processing – The IMA Volumes in Mathematics and its Applications, vol. 138, pp. 105–114. Springer, New York (2004)
35. Strapparava, C., Stock, O., Mihalcea, R.: Computational humour. In: Emotion-oriented Systems, pp. 609–634. Springer, Berlin (2011)
36. Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavvaf, N., Fox, E.A.: Natural language processing advancements by deep learning: a survey (2020). arXiv:2003.01200
37. Tur, G., Celikyilmaz, A., He, X., Hakkani-Tür, D., Deng, L.: Deep learning in conversational language understanding. Deep Learning in Natural Language Processing, pp. 23–48. Springer, Berlin (2018)
38. Tur, G., De Mori, R.: Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. Wiley, Hoboken (2011)
39. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. ACL (2002)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need (2017). arXiv:1706.03762
41. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signal processing: state-of-the-art and future perspectives of an emerging domain. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 1061–1070. ACM (2008)
42. Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: emotional conversation generation with internal and external memory. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)