Assessing the quality of natural language text data

Daniel Sonntag
DaimlerChrysler Research and Technology, Ulm Germany
daniel.sonntag@daimlerchrysler.com

Abstract: We follow an empirical approach from data quality toward text quality, where the expectations of the consumer, human or machine, take the centre stage. We try to obtain numerical text quality statements which must be interpreted for the expectations of the user and suitability for automatic natural language processing (NLP) separately. We state that apart from text accessibility today only representational text quality metrics can be derived and computed automatically. Interestingly, text quality for NLP traces back to questions of text representation.

1 Introduction and background

Data Quality has always been an issue in database systems. Correct data entries, no duplicates and referential integrity are examples of quality assets in relational databases, and quality factors such as correct process models are examples in data warehouse management, which introduced data cleaning as integral part of data maintenance processes. But even on the simple, structured data fields like customer addresses, assessing data quality measures and cleaning inconsistent data entries has proven to be a challenging problem, though a very important one [GH01]. Linguistic text data are ubiquitous and text information retrieval is a crucial component of modern information systems having a big impact on everyday people: modern information systems contain more and more unstructured data like reports, specification sheets, customer feedback emails, reports, web pages, and discussion papers. The former small market of information retrieval systems has shifted to a huge commercial natural language text processing market. Unfortunately, the need for data quality, too, which now faces more abstract and more inscrutable data in form of natural language text with complex syntax and semantics. The question we address is, which data quality requirements can be stated for digital natural language text and which quality measures can be automatically computed, against the background of many unsolved problems in natural language understanding and computational linguistics. The most serious shortcoming of today's language technology is the lack of methods that get at the real contents of texts [Us02]. Of course, the content defines some of the most important quality measures. Therefore we face many reservations on effective quality measures, nonetheless with some opportunities for closed domains, which we are going to explore. We focus on representational text quality, which can in part be computed automatically, and describe, how easily the text can be understood by a human, and for second, how successful the text can be automatically classified, retrieved, or information extracted from.

2 Text quality dimensions

We begin by a clear understanding of what data quality means to data consumers. It is the concept of *fitness of use* that emphasises a consumer viewpoint of quality because getting information from the data or text is ultimately the consumer interest; he jugdes whether the data is fit for use [Ju89]. This viewpoint provides a better perspective on natural language text, than quality improvement approaches focusing narrowly on accuracy can provide. Therefore we base our conceptual framework on the four target categories matching to data quality dimensions [WS96]. Figure 1 highlights the dimension we identify for text quality, which we will explain further.

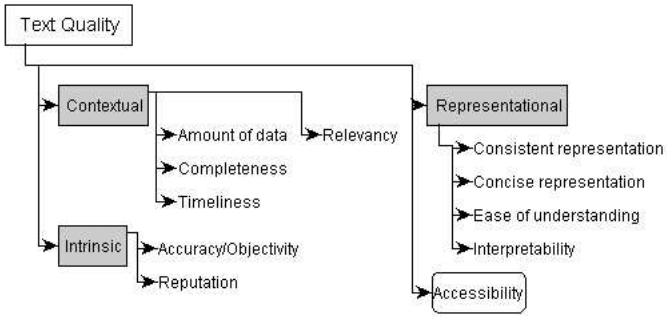

Figure 1: Text quality dimensions.

Intrinsic text quality: Intrinsic data quality is declared by accuracy (data are certified error-free, precise) and objectivity (unbiased). In [WS96] objectivity and reputation (of data and data sources) are additionally counted to strong empirical intrinsic data quality factors. This can be accepted for text data without any qualification, too. The problem is, as pointed out in the introduction, that state-of-the-art natural language processing (NLP) components lack methods to get at the real content of texts, especially the inferable knowledge contained in the text. Therefore, no comparison to other document's content is possible. By the way, semantic processing and automatic reasoning on high-level semantic meta data are also beyond current system capabilities. On the other hand, intrinsic data quality properties that can actually be computed, such as reputation, are very common for the textual domain. For instance, author-, institution-, and reference entry type of a research paper offer good means for assessing a quality reputation model. Well-known researchers from renowned institutions represent good reputation, which is even transitive to the references mentioned within the document. But reputation and similar indications should not be too widely used to let us draw conclusions from, e.g. towards accuracy and believability. This inference is indirect and not based on linguistic content. Therefore we conclude, that apart from reputation, it seems that no other intrinsic text property can be automatically modelled in our day. **Accessibility of texts:** Accessibility is obviously a characteristic of the information system. Apart from access security, it is the possibility to automatically retrieve data by a query which is a pure machine-consumer's expectation. In

traditional information retrieval applications, a set of documents is indexed and found with a text query [FBY92]. For text data, accessibility plays are more important role than for traditional numerical or string-based data. In Data Retrieval only exact matches to a query are necessary and considered, whereas in Information Retrieval documents with a certain probability of relevance to the query are searched. Information Retrieval queries are technically speaking k-nearest-neighbour queries with similarities adopted to the specific information need. This distinction is very important and reveals a typical quality aspect for machine consumers, how easily a text document can be retrieved from a database (confer precision/recall).[1] **Contextual text quality:** Text quality must be considered within the context of the task at hand. Contextual data quality is assessed on the basis what the text is used for. Since tasks and their contexts vary across time and consumers, attaining high contextual text quality is a challenge. In the common approach, appropriate contextual parameters must be set manually, and for every task individually. For instance, a weather report is meant to be short and must not necessarily contain complete sentences, or a pilot manual must be written with a restricted vocabulary. For automatic NLP, the context in which the texts are used has a different main component as in the the common approach: Fitness for use in terms of suitability for automatic processing by computers. Texts can better be processed automatically, if the text representation is suitable for automatic processing, which leads from contextual to representational quality aspects we focus on in the following. **Representational text quality and deficit resources:** Representational data quality includes aspects of data formats in form of concise and consistent representation, and meaning in form of interpretability and ease of understanding. This suggests, that texts are well-represented, if they are concise and consistently represented, but also interpretable and easy to understand. In our approach, the absence of redundancy is ranked among a consistent representation. [RD00] formulates single-source problems and multi-source problems which can be directly related to textual data and representational deficits.[2] Single source problems are *wrong formulated data values*, *typing errors*, *different spellings of same word*, *co-reference* problems, and *lexical ambiguity*. We adopt the *wrong formulated data value* problem to textual data by the following lexical word mapping problem: Directly referring named entities, like data and time expressions, can have different syntax, like *3. October, 3rd of October, October 3rd*. As *co-reference* problem, consider a text where a person is named, such as Bill. This named entity can be referenced by Mr. Jones, he, the man, or Susan's husband. The data values itself bear the *co-reference* problem and if the reference is not correctly resolved, we are using data values not referring to the same person as originally intended by the writer. For third, *lexical-ambiguity* problem means that individual words or phrases can be used in different contexts to express two or more different meanings. Multi-source problems are *homonym name conflicts* and *document duplicates*. Homonym conflicts result from a special type of lexical ambiguity in which a word having the same spelling as another (homograph) is differing from it in meaning such as the noun *bear* and the verb *bear*. Likewise, in financial magazines, *bank* means financial institution, different from a bank in a natural park.

[1] Text Retrieval is one of the most prominent automatic text processing applications. Classification is quite similar, because it also takes similarity measures into account.

[2] In databases we can additionally differentiate between schema and instance problems, which is more difficult for unstructured textual data, since the schema, if it exists at all, is often not specified in an explicit way.

3 Text quality metrics

You can't control what you cannot measure - it says in software architecture. Since we are primarily interested in computable text quality metrics, this applies in our case, too. In the operational database process, data cleaning can only be started, if metric values can automatically be obtained and interpreted. Anchoring data quality dimensions in ontological dimensions [WW96] obtains a good means how to measure text quality, though data audit guidelines and procedures are hard to express. The audit guideline says to analyse data quality based on inconformities between two views of the real world system: The view obtained by direct observation and the view inferred by reading the text. We introduce a third viewpoint, from the NLP devices to formulate effective quality metrics. Taking a closer look at the consumers, we have identified two groups for which text quality plays a major role, humans and automatic NLP. Human text quality is measured in the inner cycle in figure 2, the dotted line starting in human information processing points out, that intrinsic and contextual quality can only be measured between these viewpoints. Assessing text quality measures for both groups and relate them to each other, which is the intention, we keep in mind that quality metrics concerning the interpretation of text are too difficult to compute automatically and too pragmatic because they heavily depend on the linguistic content and the specific situation in which the text is being used. For automatic NLP, we condition the context in which the texts are used by assuming, that text is used for information needs which are apparent in the text, candid, and can be extracted by a native speaker. With this in mind, text quality metrics restrict to accessibility/representational metrics that can automatically be processed. This forms the outer cycle, whereby the human plays the supervisor and we measure the quality of automatic processing between the viewpoint of the human and the outcome of the automatic NLP. Finally, better text representation means to make text more suitable for automatic NLP, by i.e. better interpretability and less ambiguous content. For example, word sense disambiguation (WSD) is used to identify (and resolve) homonym conflicts. A quality metric would measure, how often WSD was needed (outer cycle) by automatic NLP, and how often performed successfully (inner cycle).

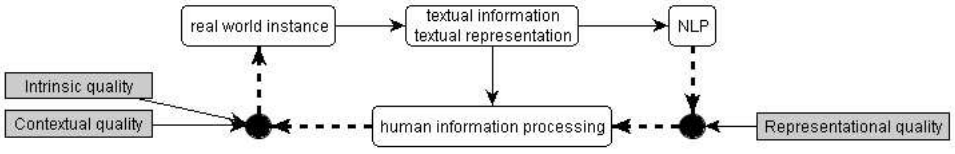

Figure 2: Text quality audit guideline.

Computational metrics: The base model is to define a normative state N^T on training set T in respect to the quality aspect in question.[3] We then derive patterns P^S from new texts S, like frequencies, covariance matrices, or other parameter loads. We then compare both

[3]The normative state can also be set by manual parameter settings, e.g. only few spelling errors.

resulting in a degree of deviation $\delta(N^T, P^S)$. The less this degree, the better the quality. Representational quality can be measured by lexical and document metrics. On the lexical level, single words used in the text are in focus, and lexical relations between the words (linguistically: wordforms). In the simple case, spelling quality can be quantified quite easily by frequency patterns. In addition, class compression models [WBMT99] can be used for automatic tagging of named entities, and Latent semantic indexing [Ho99] is a computational framework for WSD, whereby the ambiguity itself can be measured by a lookup of the synonym classes in an ontology. On the document level, the text as a whole is in focus. In this case, grammatical errors can be quantified, such as number disagreement, as in *The dogs is barking.* In order to obtain normative document parameters, e.g. Language entropy models [Ch93] can be used to assess the uncertainty of the information to be extracted.

Summary: We have defined text quality dimensions and computational metrics for representational text quality, measuring text quality for automatic NLP separately from what it means to humans. This distinction defines the basis for future confirmatory studies on text quality along the quality dimensions introduced and makes clear that text quality enhancements can only be achieved, if the consumer, human or machine, is known in advance.

References

[Ch93] Charniak, E.: *Statistical Language Learning*. Cambridge: MIT Press. 1993.

[FBY92] Frakes, W. and Baeza-Yates, R. (Eds.): *Information Retrieval Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, New Jersey. 1992.

[GH01] Grimmer, U. and Hinrichs, H.: Datenqualitätsmanagement mit Data-Mining-Unterstützung. In: *Praxis der Wirtschaftsinformatik*. dpunkt.verlag. December 2001.

[Ho99] Hofmann, T.: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*. pp. 50–57. Berkeley, California. August 1999.

[Ju89] Juran, J. M.: *On Leadership For Quality: An Executive Handbook*. New York, N.Y: The Free Press, A Division of MacMillan Inc. 1989.

[RD00] Rahm, E. and Do, H.: Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering, Vol. 23 No. 4*. 2000.

[Us02] Uszkoreit, H.: New chances for deep linguistic processing. In: *Proceedings of COLING'02*. Morgan Kaufmann Press. 2002.

[WBMT99] Witten, I. H., Bray, Z., Mahoui, M., and Teahan, W. J.: Text mining: A new frontier for lossless compression. In: *Data Compression Conference*. pp. 198–207. 1999.

[WS96] Wang, R. and Strong, D.: Beyond accuracy: What data quality means to data consumers. *Journal of Management of Information Systems, 12, 4*. pp. 5–33. 1996.

[WW96] Wand, Y. and Wang, R. Y.: Anchoring data quality dimensions in ontological foundations. In: *Commun. ACM 39,11*. pp. 86–95. 1996.