Rule-based approach to text generation in natural language - Automated Text Markup Language (ATML3)

Armin Bauer[1], Niki Hoedoro[1], Adela Schneider[1]

[1]Aexea GmbH, Stuttgart, Germany

Abstract. The need for text online is quite large. The majority of websites still make use of text as their main representative form for information. Most people can easily comprehend issues in text form and often prefer this type information for communication. Still the prices for written texts have been falling over the last years. We present a system that can generate natural language text in multiple different languages, in large quantities and at low costs due to the utilization of multiple layers of rule based reasoning.

Keywords: natural language · rule · rule based · grammar · AX · text · NLG · Linguistics

1 Business Case: Why Automatic Text Generation?

We at AX-Semantics have created a software solution we use to generate texts. Our business model with this software is selling texts to our customers. The generated texts are of different types, for example product descriptions, journalistic texts about repeating events or functional texts, for example in the finance sector. Many other kinds of text are possible.

1.1 Higher Demand for Text Online

The need for text online is quite large. A majority of websites still make use of text as their main representative form for information. Most people can easily comprehend issues in text form and often prefer this type information communication. The text-heavy nature of the Internet is also due to the text-centered function of search engines. Web crawlers interpret text, but for images, videos, graphics they are essentially “blind”; the information contained in these forms of content cannot be directly extracted. Web crawlers are dependent on supplementary text during the analysis of content. Only recently has there been an increase in significant solutions for text-independent image recognition, for instance Google’s approach via Convolutional Neural Networks.

Simultaneously, larger search engines are setting the standards for text quality: Only those who comply with the requirements for keyword application and text structure stand a chance to achieve a good rank on a search engine’s results list.

Google’s penalty for “duplicate content” also maintains the demand for large text quantities, because website operators are supposed to present unique content on their websites. Thus they cannot simply release the manufacturer’s texts for e-commerce product descriptions, but rather write the product texts for their clients themselves. To illustrate the required text quantity: a medium sized online shop has approximately 200 000 available products. The numbers in the category “electronics” are higher, as a result “Conrad GmbH’s” offers approximately 600 000 products in its online shop.

1.2 Text Production Costs Increase Disproportionately to Text Quantity

The cost of text production depends fundamentally on the type of text, text length, the research costs and text quality. However it increases by a substantial amount for similar texts disproportionately, because of the extreme growth of administrative expenses. With the compliance of quality standards, the biggest problem to arise is that the speed of text production is not easily scalable. Manually producing large text quantities takes time, as every text has to be individually written. An author can only write a certain amount of text in a certain amount of time.

Agencies from the “Content Broker” sector have attempted to solve this problem by establishing large pools of authors. However, the authors’ varying levels of writing skills and the broad instruction and briefing for individual projects mean that the client has to accept quality fluctuations and relatively high prices.

To comply with these requirements, a software solution is the obvious answer: AX Semantics is able to quickly produce large quantities of text at consistent standards of quality.

The normal way of generating texts is obviously using editors to write them. This is however limited because qualified editors are usually not available in large quantities and their writing styles differ. Using a software solution that can be trained to match a writing style, it is possible to generate hundreds of thousands of texts in the same quality and writing style. With our system this happens in a few hours and in different languages. Human editors and translators would need months or even years to do that.

1.3 Automated Text Generation Fulfills Specific Online Requirements

Another characteristic of the Internet, that increases the need for text, is the current localization and personalization trend. Internet users expect information that is tailored to their specific situations: they read weather reports for their own city, are interested in news about their region and want to know how their favorite sports star did in the last game or race. Media corporations have to adjust to these new demands, though the production costs are also high, and with the conventional methods of production, are normally unprofitable.

The software can thus offer support to media corporations and also fulfill the demands of smaller target groups. With AX one can compile specialized articles for a sports magazine for Bayern fans or even specifically fans of Philip Lahm without excessive costs. It is even possible, in principle, that communications and the depth of detail can be adjusted to the demands of an individual reader, if these were previously specified as important. Until now a Philip Lahm fan had to gather information and details from marketing-orientated websites for the player or team or from fan sites and forums. This is an additional benefit, which had not existed in the professional journalistic sector until now, that the reader may also be willing to pay for.

1.4 Examples of the Range of Application of Automated Text Generation

Where usually a trained writer would be needed, a software solution can be used to generate different text-types for different needs. Thus different kinds of business cases for different kinds of customers arise.

Example 1: Online Shops and Producing Companies
Studies prove that, for various reasons, well-written texts in online shops strongly increase the probability of a product being sold. A company producing products is required to have a product description in the datasheet that comes with the product.
If a text for describing every product on Amazon.de in Germany was generated, that would make over 200 million texts.

Example 2: Stock Reports
When reading a financial report one receives the numbers, which it is based on in a strongly conditioned form. One receives, for example, not only the current share prices at different hours, but also highly compressed information such as “increase of 50 points at 8:00 up to 120 points at 11:00”. This then includes a benefit in that the reader has to otherwise interpret this by themselves through mere observation of the numbers.

Example 3: Financial Reports
Every medium to large company is required to report about their business at least 4 times a year. In Germany there are about 12.000 stock based businesses.

Example 4: Personalized News
Many users of press products have the expectation that their interests will be considered when articles are written. For instance, it is easy to see that in a sports report, the accomplishments of the reader’s favorite player are specifically catered to, so that a stronger commitment is achieved between readers and an online newspaper.

2 Technological Challenges

2.1 Structured Data as a Requirement for Automated Text Generation

The basic requirement for data used in text generation is that the data set containing the descriptive data is made uniform and standardized. For example, a numerical data field cannot use the digit “1” at one point and then the character string “one” at another.
The data should also of course be consistent in structure and always contain the same field names.

2.2 Implementing Multiple Languages

Grammars and grammatical structures between different languages greatly differ. In a later part of this paper we will see how different grammatical properties between phrases in a

sentence are inherited and enforced through other sentence parts. This principle is called government (case government) and it works differently in various languages.
Languages can however be expressed through different rule sets describing their syntax. The technical challenge is to build a machine that can be used with different languages and is flexible enough to express sentences in multiple languages.

2.3 From Semantics to Grammar and Syntax

In some languages there are agreements between syntax and semantics. An example is that a group of things that does not consist of exactly one element in German is causing plural form or that talking about some event that has already passed (giving the date) is forcing a past tense.
In some languages there are much more complex rules to that than in English. They don't only know plural and singular but have more different grammatical numbers or treat groups of things different than groups of groups of things.
Those properties arise as agreements that aren't a property of syntax only but are linked to the semantics of a sentence. Still getting them wrong is considered an error in grammar.

3 Overview: How does the Software Work?

The software is based on a complex algorithm that is capable of connecting data from different databases and, from that, identifies relationships, correlations, influences and special features. A system of rules, which builds the basis for a conclusion, is designed for every type of news. The content is then integrated into a narrative framework and implemented in natural language. AX's distinctive feature is that the software uses generative grammar and can utilize syntactic rules and inflection. The semantic model developed from the analyses is in this way linguistically reproduced and refined. The results are reader-orientated, comprehensible and engaging texts.

3.1 Reasoning: From Data to Conclusion

One can assume that, in e-commerce, the products' text in online shops takes the role of a salesman in a store. Therefore it has to deliver all the information to the customer that a salesman would impart with experience and world knowledge.
For the text to be perceived as sufficiently informative and useful, it must also include more information other than only the facts from a product specification sheet. This can be, for instance, a scope of application and particular strengths of the product. Here rule-based reasoning is mostly used to create extended information retrieval and context for a product text.
A typical application for this kind of information retrieval would be, for example, the comparison of a product with another well-known product. One can ascertain, that a cellphone with a lightweight is suited particularly to users, who want a device with most

mobility possible. This of course functions exactly the same way for other characteristics, such as standby time or talk time before the battery is empty.

In addition, it is even possible to execute an evaluation throughout a database of all products, to see if there is a device on hand, that is superior in a particular characteristic.

An additional example of usage for rule-based reasoning would be to derive specific uses from a product's properties. For example, one could assume that a wristwatch with splash-proof protection is in fact suitable for sport, but not for swimming or diving.

The machine is trained to write particular sentences and in this way the use of rule-based reasoning offers an advantage; all possibilities of the characteristics one wants to describe can be precisely and reproducibly programmed.

It is important that a direct recommendation or warning can be reliably given and not erroneously via statistical noise or missing information (e.g. The "spray water protection" sentence does not occur because of statistical inaccuracies in machine-learning).

It is possible, however, to extensively automate the generation of such rules.

The focus in ATML3 lies on the production of texts, from which the generating machine chooses a language form that was supplied by an editor. This way, one can use the same machine, but different training programs to generate, for example, press texts in different tones, product texts with stronger or subtler sales tactics or neutrally formulated factual texts.

3.2 Conception to Text: The Construction of Grammatical Benefits with ATML3

Another layer that the automatic text generation rule-based approach can be used on is linguistic deduction in sentences. With sentences especially in the German language it is common that words correlate with one another and thus define their inflexion. At the same time, the words in a German sentence obtain parts of their grammatical characteristics from other words and other parts from semantic agreement.

Der fleißige Student fällt durch die schwierige Prüfung. (1)
(The diligent student fails the difficult exam.)

In this simple example sentence there are already several grammatical references that impact the linguistic characteristics of the words. Some examples are:

- The nominal phrase *Der fleißige Student* (the subject of the sentence) determines that the verb *durchfallen* is in singular.
- The verb *durchfallen* in the third person singular determines that the nominal phrase *die schwierige Prüfung* is in accusative case.
- Since the article *die* is positioned before the noun *Prüfung*, it inflects the adjective *schwierig* to *schwierige*.

One can easily see that words in the sentence contain many connections with one another in order to compose a well-formed statement. In more complicated sentences these connections become far more complex but can still be modeled using ATML3.

Moreover, different languages have even further varying methods to display these relationships between the parts of speech within a sentence. For example, in the Spanish language most adjectives are placed after the noun they describe.

A complex system of rules, with which the linguistic components in sentences can be produced, is also formed on this layer.

Grammatical properties in ATML3 are relatively limited because there is a given set of rules exclusive to the target language. Our example sentence in ATML3 would be represented as follows:

[student, adj=yes, id=subject] [G:verb=durchfallen,
grammar-from=subject, id=verb] [prüfung, adj=yes, (2)
grammar-from=verb#X2]

Here the reference to all sentence components is denoted by other sentence components respectively. Because there are verbs that assume more than one object, each verb's particular case is represented with #X2.

Within the text machine's linguistic functions, the characteristics the verb will obtain from the subject (or characteristics a verb will obtain from a noun) will be selected and with that the deduction engine arrives at the conclusion that the verb can obtain number (in our case singular).

Ultimately, the verb can be rendered alone (third person singular → *fällt durch*) and also specify the reference *die schwierige Prüfung* to the argument case X2 for the verb *durchfallen*. Following that, the sentence is completed.

If one changes the sentence a little, for example in:

[*appeal:self id=subject] [G:verb=durchfallen,*
grammar-from=subject, id=verb] [prüfung, (3)
adj=yes, grammar-from=verb#X2]

The *appeal:self*-container would determine the subject *ich* in the sentence and could therefore dictate the verb to be first person singular. From this, the resulting sentence is *Ich falle durch die schwierige Prüfung.*

4 Results and Conclusion

Our system has generated millions of texts to date. Among them are mostly product descriptions and journalistic articles.

The rule-based approach is very well suited to be used in natural language generation and solves the problems at hand reliably and fast.

It is especially fascinating that there are many varying sub-systems with different functions, which, despite their completely different fields of application, engage one another and gradually solve the complex task of "text creation".

5 Importance and Impact

As mentioned in chapter 1, the need for text and especially for personalized text is increasing. Thus natural language generation is becoming an important feature for many businesses in e-commerce, reporting and journalism.

The ATML3 language is currently being made available to everybody. This means that we have a beta-program running that allows interested developers to create their own natural language generation applications using our platform and technology.

Our solution is the only one known to us to date, that allows everyone, organizations, enterprises or individuals, to turn their data into natural language that can be understood easily by their customers, stakeholders or just any interested person.

Thus this technology helps to spread information and make it easier to comprehend and use it.