# Natural Language Processing with PoolParty

Version: May 2019

## Table of Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to PoolParty</td>
<td>2</td>
</tr>
<tr>
<td>Resolving Language Problems</td>
<td>3</td>
</tr>
<tr>
<td><strong>Key Features</strong></td>
<td></td>
</tr>
<tr>
<td>Entity Extraction &amp; Term Extraction based on Knowledge Graphs</td>
<td>4</td>
</tr>
<tr>
<td>Shadow Concepts</td>
<td>5</td>
</tr>
<tr>
<td>Word Sense Disambiguation</td>
<td>6</td>
</tr>
<tr>
<td>Entity Extraction based on Machine Learning</td>
<td>6</td>
</tr>
<tr>
<td>Relation Extraction</td>
<td>6</td>
</tr>
<tr>
<td>Document Classification</td>
<td>7</td>
</tr>
<tr>
<td>Semantic Similarity and Content Recommendation</td>
<td>7</td>
</tr>
<tr>
<td>Corpus Learning</td>
<td>8</td>
</tr>
<tr>
<td>Semantic Knowledge Graphs</td>
<td>8</td>
</tr>
<tr>
<td><strong>Some NLP-based Applications</strong></td>
<td>10</td>
</tr>
<tr>
<td><strong>Roadmap</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>Contact</strong></td>
<td>11</td>
</tr>
</tbody>
</table>
Introduction to PoolParty

PoolParty Semantic Suite is a globally renowned, best-of-breed semantic middleware initially released in 2009, which is available as software as a service (cloud service) or as an enterprise software to be installed on-premise. The platform provides a broad range of functionalities and APIs, from taxonomy management through text mining, natural language processing, to data integration based on linked data principles. PoolParty’s data model is fully compliant with W3C Semantic Web Standards, providing its customers with competitive advantages through superior reusability of semantic knowledge models.

In this paper we will focus on PoolParty's Text Mining & Natural Language Processing capabilities. In the core of each application that builds upon a semantic information architecture, we clearly distinguish between content & data layer, metadata layer, semantic layer, and the application with its own navigation and business logic on top.

In many cases, the least obvious distinction we make is between metadata and semantic data. To illustrate the necessity for this, let’s assume we process a document that contains entities like ‘The Pope’ and ‘Rome’. This kind of semantic metadata remains meaningless in the sense that it’s not actionable until both entities become things (or resources) that can be linked to each other or to something else like a knowledge base, to rules that can be used for more precise entity extraction, or to automatic quality checks. By adding a semantic layer that contains facts

---

1 KMWorld listed PoolParty Semantic Suite as Trend-Setting Product 2015, 2016, 2017 and 2018
3 W3C Semantic Web
like ‘The Pope is the head of state of Vatican City, which is a sovereign city-state that is entirely enclosed within the Italian capital city of Rome’ on top of simple metadata, one can quickly find the document when searching for ‘Vatican’, even when it doesn’t mention the Vatican explicitly. By adding a semantic layer/knowledge graph on top of a metadata layer, metadata becomes meaningful because it is then put in a richer context, and it becomes actionable because metadata provides no longer just strings but machine-addressable and -processable things or entities.

PoolParty’s Natural Language Processing is part of a methodology that makes unstructured and often ambiguous content processable while linking automatically extracted entities to the corresponding resources in a well-defined semantic layer. A semantic layer is a network (or graph) of things including its relations and attributes such as its various names. This layer serves like a glue to link all the information available for a certain business object (‘thing’ or ‘resource’) scattered across various repositories and data silos in order to create a complete picture of it.

Resolving Language Problems

Human beings understand in most cases what’s meant by words that can have several meanings, but how can we teach machines the semantics of ambiguous words and phrases? To exemplify a case of ambiguity, let’s take a look at a word like Jaguar. We all know how to disambiguate it because of the context in which the word is embedded. When someone says ‘Jaguar is owned by Tata Motors’, it’s clear that it’s the car and not the cat. In this sense Jaguar is a so-called homograph, a word that is written in the same way but has different meanings. Whereas homographs have to be disambiguated, synonyms refer to the same thing. One and the same thing can have various names or labels. For example, Panthera onca is the synonym of Jaguar in case it refers to a big cat. Synonyms and homographs are fundamentally different
concepts in linguistics, and both phenomena have to be taken into account by any text mining technology to increase accuracy and precision of correctly extracted entities.

In addition, polyhierarchies are frequently confused with the linguistic features from above. A polyhierarchy describes an entity or concept as a child concept of at least two parent concepts. An example of a polyhierarchy is that Jaguar is not only a Cat, but Jaguar is also a Keystone species.

Key Features

Entity Extraction & Term Extraction based on Knowledge Graphs

**PoolParty Entity Extractor** (PPX) provides a performant and clusterable semantic service that is able to extract automatically the most relevant entities and terms from a given document or text fragment. Entity extraction in this context also means that entities found in a text are automatically linked to a given resource from a knowledge graph, in which additional facts about the extracted entities could be stored and used for deep text analytics (‘knowledge extraction’).

The service is based on sophisticated algorithms that use several components of a knowledge model:

- Taxonomies based on the SKOS standard
- Ontologies based on RDF Schema or OWL
- Word form dictionaries
- Blacklists and stop word lists
- Disambiguation settings
- Domain-specific reference document corpora
- Statistical language models

Whereas extracted entities are already part of the knowledge model that is used as the basis to build the extraction service, PoolParty will also extract terms that are not yet part of the semantic
model, but scored as relevant in a given text. Such terms could be single words (unigrams) or also multi-word phrases (n-grams). PPX can handle various file formats as Word, PDF or HTML, and can be integrated in any content workflow to support semi-automatic tagging, semantic enrichment or semantic indexing. Several PPX services that analyze different aspects of a given domain can be executed in a row. Algorithms are language-agnostic and work with any major European, and also partly with Asian languages.

From a more technical perspective, PoolParty’s extraction engine is organised as units with specific functionalities that build on the output of previous units and add specific results to the output stream. The main functionalities are realised in the Term Extractor and the Term Matcher. The Term Extractor detects specific pieces of the text that are characterised as potential term candidates. The Term Matcher matches the candidates to the thesaurus model and resolves conflicting matches. The Pre- and Post-Processors prepare the text for processing and clean up the results before generating the final output.

Shadow Concepts

A common phenomenon in many knowledge domains is that the “aboutness” of documents is not surfaced, thus particular subjects are touched without mentioning those explicitly. With PoolParty’s Shadow Concept feature, hidden entities can be extracted even if not found in the analysed text. In this example, in addition to explicitly used concepts and terms, Machu Picchu is extracted from the article as a Shadow Concept. As a prerequisite, one has first to provide and analyze a representative reference text corpus to make use of co-occurrences between concepts and terms.
Word Sense Disambiguation

Word Sense Disambiguation (WSD) is a common problem in NLP. The same word or phrase can mean different things (polysemy). With PoolParty, knowledge engineers can determine path patterns of a knowledge graph that should be used to calculate a contextual model around a concept. This information will then be compared with the surroundings of the potentially ambiguous extracted entities in a given text. This approach helps to resolve ambiguities with high precision, and there is no need to define complex extraction rules.

Entity Extraction based on Machine Learning

To complement PoolParty’s graph-based entity extraction services, also machine-learning based extractors can be trained for all types of entities. PoolParty is shipped with pre-trained extractors (organisations, locations, people) but can be extended by customised extractors for any type of entity (for example brand names, proteins, or names for music bands).

Relation Extraction

With PoolParty technologies, text analytics can be brought to the next level: In order to identify specific text fragments from larger documents, several extraction services can be combined. Typically graph-based knowledge extraction, ML-based entity extraction, and also extraction of regular expressions are used together and are executed against rules expressed via SHACL⁴. Based on this approach, complex constraints and relevant relations between entities can be formulated and can be used to extract, for instance, ‘critical’ paragraphs from a contract.

⁴ Shapes Constraint Language (SHACL) - https://www.w3.org/TR/shacl/
Document Classification

For document classification, PoolParty combines machine learning algorithms such as Deep Learning or Support Vector Machines (SVM) with semantic knowledge models to classify documents automatically with high precision. PoolParty_Semantic Classifier learns from a set of pre-classified documents by extracting feature vectors based on terms and concepts stemming from domain-specific vocabularies. With this mixed approach in place, F1 scores\(^5\) of the resulting document classifier can be improved by 1-3% compared to classifiers which rely solely on unstructured text.

In addition to the Classifier module, PoolParty can also help to categorize documents, based on a well-defined taxonomy. To this end, the taxonomic structure is exploited down to its deepest level. The categorization service produces a list of categories based on the concepts (entities) extracted from a document. Matched thesaurus concepts are backtraced to their related top concepts (via skos:broader relationships), and the scores of all matching leaf concepts are integrated to a compound score. Based on this methodology, domain-specific sentiment analyzers can be developed as well.

Semantic Similarity and Content Recommendation

Finding documents or other data objects similar or related to a given one is a highly desired feature to support content authoring, information retrieval, or configuration tasks. PoolParty provides various similarity algorithms and services that exploit semantic knowledge graphs and/or extracted features of a document. As an example, users benefit from a ‘Show similar documents’ functionality while searching for relevant documents. PoolParty provides Semantic Similarity services to be included in any analytics or search application.

\(^5\) F1 score: [https://en.wikipedia.org/wiki/F1_score](https://en.wikipedia.org/wiki/F1_score)
In addition to recommender systems based on ‘similarity’, PoolParty supports the creation of recommender systems that are built upon matching rules. Such rules are part of the semantic knowledge model and help to identify objects of different types that fit with each other, or even complement each other. By that, also sophisticated ‘configuration problems’ can be solved. On the left an example illustrates how harmonising wine and cheese pairs can be found based on PoolParty’s recommender system. As an input, only descriptions of the products are needed. Relevant concepts get extracted automatically from text and are executed based on the matching rules to identify well tasting pairs automatically.

Corpus Learning

PoolParty’s text mining and knowledge modelling approach combines controlled vocabularies with statistical language models (co-occurrences for further context analysis), reference text corpora, and word form dictionaries. As a result, an extensive semantic graph of entities, terms and relations will be created partly manually, and partly automatically. A core feature in the process of creating and extending the knowledge model is based on semi-supervised learning from reference text corpora.

Based on machine learning, PoolParty extracts candidate terms and suggests automatically how to include them into the existing knowledge graph. This feature also helps to make sure that the resulting model will cover the scope of a given knowledge domain to the greatest possible degree. Corpus learning is also a practical solution to support taxonomists and subject matter experts in collaborating more efficiently.

Reference corpora can be used for various purposes besides taxonomy extension, e.g. for a more accurate scoring of extracted terms, or also for the extraction of Shadow Concepts.

Semantic Knowledge Graphs

As a core principle, PoolParty’s NLP capabilities make use of taxonomies, thesauri, and ontologies, that can partly be derived from and/or mapped to linked data sources like DBpedia, Wikidata, Geonames, etc. PoolParty provides extensive vocabulary management facilities including Quality Management, Workflow Engine, or a SPARQL query engine. As a result, PoolParty is widely used to create and maintain Semantic Knowledge Graphs on a large scale.
PoolParty makes use of Graph-based standards recommended by the World Wide Web Consortium (W3C). Users benefit from a highly agile approach for data management that will grow flexibly over time. The use of controlled vocabularies as a basis for NLP tasks makes black boxes obsolete, as well as it helps to overcome complex modelling tasks occurring with rules-based NLP services. Instead of defining complex sets of rules that define which concepts have to be extracted under which circumstances, a semantic knowledge model is used to build entity extraction services on top.

Knowledge models that are based on graphs work in a similar way to our brains, and can more easily be maintained by subject matter experts than models based on proprietary technologies used for knowledge representation.

Knowledge graphs make the meaning of concepts and words explicitly available through the usage of Semantic Web standards. The graph-based nature of such semantic models is expressive enough that even hidden relations can be used for highly precise text classification and annotation.

6 IDC White Paper: How Semantic Technologies Enable Domain Experts to Steer Cognitive Applications
Some NLP-based Applications

**Semantic Search**
PoolParty's Graph Search Server provides a full-blown semantic search interface that integrates with MarkLogic, Elastic, Solr, or with several graph databases as Stardog, GraphDB, Neptune, AllegroGraph or Virtuoso.

**Question Answering**
Intelligent help desk systems benefit from expressive semantic knowledge models that support users with guidance and a more precise interpretation of natural language queries.

**Recommender Systems**
PoolParty's ability to roll out a semantic layer builds the basis for precise recommendation services and personalised user experience that go beyond simple Vector Space Models and TFIDF.

**Linked Data & Data Integration**
PoolParty's graph-based data model in combination with its text mining technologies provide a robust framework for data integration efforts dealing with structured and unstructured information.
Roadmap

PoolParty’s NLP and text mining capabilities are continuously developed based on insights from deep market analysis, customer requirements, and a broad range of research and innovation projects.

Features and topics to be highlighted are:

- Fact extraction: based on machine learning methods and PoolParty’s ontology module, facts like ‘Person X starts working at Organisation 123’ will be extracted and transformed into RDF triples. First prototypical implementations have been completed successfully.
- Extended support for Chinese: Concept extraction based on controlled vocabularies works already for Chinese in the current version. Corpus analysis including the extraction of free terms and candidate terms will be supported in an upcoming release as well.
- Extended graph-based similarity: The current version of PoolParty already makes use of knowledge graphs to identify similarities between documents and data objects in general. Nevertheless, enhancements on the precision and configurability of the service will continuously be added in the upcoming releases.
- Topic modelling: Deep analysis of document corpora based on topic modelling helps to derive relations between concepts and terms even more precisely.

For further details, please refer to the comprehensive overview of PoolParty releases from recent years, and a list of research papers that relate to PoolParty software.

Contact

Andreas Blumauer
CEO, Semantic Web Company

- PoolParty Semantic Suite
- LinkedIn
- Twitter
- YouTube

7 https://semantic-web.com/research-projects/