Advanced Linguistic Search

ALiSe White Paper

August 2015

1. Table of Contents

2. Introduction

ALiSe is a textual search engine with advanced linguistic support. It combines the textual search with metadata information. The textual search is influenced by several linguistic rules which exploit different aspects of linguistic similarity. ALiSe is built on SOLR-LUCENE

ALiSe serves a query by returning validated results, each with a linguistic score and detailed matching information.

ALiSe exposes several textual search parameters which can be used for advanced search configurations.

ALiSe is designed for searching textual expressions, but it can also be used for document search/matching based on built-in logic of combining expression-level results.

ALiSe supports automatic indexing of the reference data (textual expressions & active metadata fields). It also offers the possibility of full/selective re-indexing in order to reflect the latest changes to the reference data.

Enabling a language in ALiSe requires only the availability of linguistic resources for that language.

ALiSe supports regular expressions and can also be utilized with Boolean operators & metadata joins creating an unlimited range of search possibilities. In case of Boolean or join searches, the score might reflect business logic rather than linguistic similarity.

3. Key Terminology

- **Content Word**
  A meaning-carrying word or phrase: nouns, most verbs, adjectives, adverbs, for instance *personal computer*, *lumière*, or *Drogenkartell*
- **Functional Word**
  Words with little lexical meaning that help to express grammatical relationships: typically articles, prepositions, conjunctions etc.
- **Query expression**
  The instruction/request the user/application is submitting to the search engine for look up
  - **Reference expression**
    The retrieved text string (part of the text corpus that is being searched through) that matches the query expression

4. Search Parameters

4.1. Basic Search Parameters

4.1.1. Search Type

Search Type refers to the mode of matching the content words of the query versus the reference expressions.

- **Full Coverage Plus (FC+):**
  All query and all reference content words need to match
- **Full Coverage (FC):**
  All query content words need to match
- **Full Coverage Reverse (FCR)[1]:**
  All reference content words need to match
- **Partial Coverage (PC):**
  Some query & expression content words need to match
- **Fuzzy Match (FM):**
  Similar to partial coverage but unmatched reference content words get penalized (as reflected in the linguistic score)

Below table compares the effect of the selection of the ALiSe Search Type. The reference expression is here a contract title (content words underlined):

Charter of the United Nations

Query	FC+	FC	FCR	PC	FM
Charter United Nations	+	+	+	+	+
United Nations	-	+	-	+	+
Release of the Charter of the United Nations	-	-	+	+	+
UN Charter	-	-	-	+	+

[1] This is close to what is known as *terminology recognition* in translation memory (and authoring) workflows: the reference set is then a terminology database, and the query expression is the to-be-translated (resp. the to-be-terminology-checked) sentence.

4.1.2. Linguistic Rules

- **Character Capitalization**

  Tolerated in word matching or not

Query	Reference	Tolerated: Case insensitive	Not tolerated: Case sensitive
Resistant Bacteria	Resistant bacteria	+	-
	Resistant Bacteria	+	+

- **Character Canonization**

  Tolerated in word matching or not

Query	Reference	Tolerated: Orthographical variants do not matter	Not tolerated: Variants matter
Bafög-Erlaß	Bafoeg-Erlaß	+	-

- **Punctuation**

  Allowed to vary or not - according to the mode of punctuation assignment to content words as well as the punctuation strictness requirements.

Query	Reference	Tolerated: Punctuation does not matter	Not tolerated: Punctuation matters
Race Carver Twin Rocker	Race Carver, Twin Rocker	+	-

- **Functional Words & Phrases**

  Determines whether functional words & phrases are to be distinguished from content words. If yes, they will be treated according to the mode of Functional Word/Phrase assignment to Content words as well as the Functional Word/Phrase strictness requirements

Query	Reference	Tolerated	Not tolerated
Charter United Nations	Charter of the United Nations	+	-

- **Word Morphology**

  Determines whether morphological variations are tolerated in word matching or not. ALiSe distinguishes seven configurable incremental levels of word morphological variation (ranging from stem character similarity to identical words)

Query	Reference	Tolerated: Any allomorphs	Not tolerated: Exact form required
approved version	approved versions	+	-

Synonym Words & Phrases

Determines whether the text matching allows for word/phrase synonyms or not

Query	Reference	Tolerated: Synonym expansion	Not tolerated
approved version	ratified edition	+	-

Taxonomical Proximity

Determines whether the text matching allows for word taxonomical proximity (according to associated distance thresholds) or not. It is a loose form of "synonym" expansion – in order to restrict it to the appropriate sense of the query words, word sense disambiguation is applied (relying on the taxonomical information)

Query	Reference	Tolerated: Semantic close "synonyms"	Not tolerated
approved paper	approved publication	+	-

Compounding

Determines whether the Content words can be compounded or not – several parameters determine the compounding handling (partial/full compounds, size and number of compound parts, etc.)

Query	Reference	Tolerated: Individual Constituents match	Not tolerated: Constituents are not sufficient
Schwangerschaft Test	Schwangerschaftstest	+	-

Spelling

Determines whether spelling variants (as well as misspellings) are tolerated in word matching or not – the tolerable number of spelling variants can also be determined

Query	Reference	Tolerated	Not tolerated
appproved	approved	+	-
approved	approvd	+	-

- **Word Order**

  Determines whether different order of matched content words is tolerated or not.

Query	Reference	Tolerated: Word sequence not relevant	Not tolerated: Word sequence relevant
United Nations Charter	Charter of the United Nations	+	-

- **Negation Words & Phrases**

  Determines whether negation words & phrases are to be distinguished from functional words/phrases. If yes, they will be treated according to the mode of negation word/phrase assignment to content words as well as the negation word/phrase strictness requirements

Query	Reference	Tolerated: Opposite sense allowed	Not tolerated: Opposite sense not allowed
delay	without delay	+	-

4.1.3. Quality Thresholds

Linguistic Score Threshold

Sets a minimum required quality of matching – the linguistic score formula includes several configurable parameters (weighting factors for all linguistic rules)

4.2. Advanced Search Parameters

Below illustrates most relevant advanced search parameters that influence the result. Yet, more ALiSe configurable parameters exist. Some could also be exposed through the API (based on Use Case) while others are, clearly, System (Administrative) parameters that affect the performance of the algorithms employed throughout ALiSe (for example, performance sensitive settings for algorithm optimality).

4.2.1. Query & Result Content Word Gap Restrictions

The following parameters can be configured (where applicable according to ALiSe Search Type):

- **Maximum Query Gap**
  maximum tolerable number of uncovered query consecutive words
- **Maximum Reference Gap**
  maximum tolerable number of uncovered reference consecutive words

4.2.2. Rules for Attaching Functional-Words & Punctuation to Content Words

The matching of Functional-Words/Punctuation can be driven (or not) by the Content words to which they are attached. The following modes are supported for Functional-words & Punctuation (independently).

- **Left/Right**
  Functional-word/punctuation attached to the next/previous content word (or the last/first content word in case of unavailability of next/previous content word)
- **All**
  Functional-word/punctuation not attached to Content word – to be matched independently of content word matching
- **External**
  The attachment of the functional-word/punctuation does not have a specific direction - it is defined on the corresponding linguistic resources per case
- **Dynamic**
  The attachment of the functional-word/punctuation does not have specific direction. It is decided based on what best serves the actual content word matchings.

4.2.3. Strict Level & Scope for Functional-Words & Punctuation

Functional-words & Punctuation (together or independently) can match at four levels of incremental quality:

- **OFF**
  the Functional-words/Punctuation are allowed to mismatch without restrictions
- **Compatible**
  the Functional-words/Punctuation belong to the same compatibility group
- **Same**
  the Functional-words/Punctuation match in any order
- **Identical**
  the Functional-words/Punctuation match in left-to-right order

Whatever the strict setting is (not OFF) it will be imposed based on the corresponding scope setting:

- **Local**
  strict setting applied on the Functional-words/Punctuation as attached to each content word/phrase pairing

- **Global**
  strict setting applied on the Functional-words/Punctuation regardless of their attachment to content words

4.2.4. Parts of Word Matching Options

ALiSe can match complete words and/or their parts (compounds)

Partial/Complete
Determines whether partial compound matching is tolerated (Partial) or not (Complete)

Query	Reference	Partial	Complete
Innenministerium	Bundesinnenministerium	+	-
Bundesinnenministerium	Bundesinnenministerium	+	+
Bundesministerium	Bundesinnenministerium	+	-

4.2.5. Start & End Restrictions

Determines whether the retrieved results need to satisfy (or not) Result Start and/or End matching restriction.

Query	Reference	Start ON	Start OFF
Bundesministerium	Bundesinnenministerium	-	(+)
	Bundesministerium des Inneren	+	+

Query	Reference	End ON	End OFF
Innenministerium	Bundesinnenministerium	+	+
	Innenministeriums-Erlaß	-	+

5. ALiSe Architecture & Deployment

ALiSe is built on top of Solr/Lucene (supports both 3.x and 4.x versions) though Solr/Lucene's plugin infrastructure and provides seamless integration with existing Solr installations, or can be used as a standalone installation. The figure below depicts the overall ALiSe system architecture.

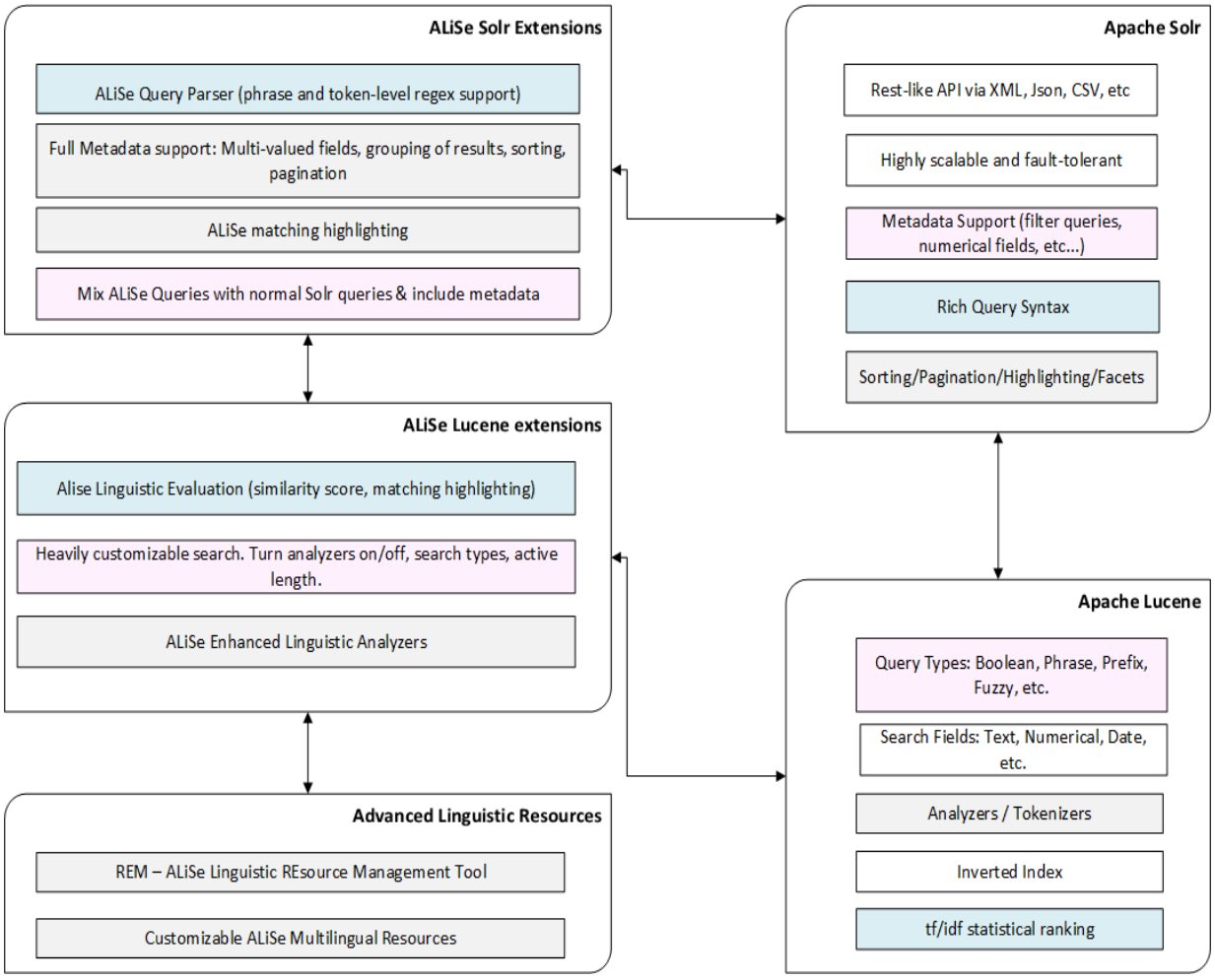

Figure 1: ALiSe Architecture - The Blue/Pink/Gray colors indicate the places where ALiSe plugs in and extends the Solr/Lucene components

ALiSe *Linguistic Evaluation* Algorithm

- *ALiSe Linguistic Evaluation Algorithm* - Provides a custom linguistic score between the query expression and each result. Supports full and partial matches and provides fine-grained configuration.
- *Regular expression support* - Phrase and token-level regex support
- *REM* - ALiSe Linguistic REsource Management web suite

ALiSe Solr Extensions

- *ALiSe Query Parser* - It can be registered as a custom component in solrconfig.xml and provides Solr integration via the normal Solr query interface.
- *Metadata support* – Multi-valued fields, grouping and sorting of results, pagination.
- *Two-way Matching Highlighting* - ALiSe provides detailed highlight of each match on both the query and the result.
- *Query Types* – Mix ALiSe Queries with Boolean Queries, Phrase Queries etc. and include metadata

ALiSe Lucene Extensions & Advanced Linguistic Resources

- *ALiSeQuery* - ALiseQuery is a custom query type that can be used as any other Lucene query such as BooleanQuery, PhraseQuery etc. It integrates our Linguistic Evaluation score into Lucene/Solr.
- *ALiSeSimilarity* - We provide out custom Lucene Similiarity class specially tuned for better ranking of the results.
- *ALiSe Enhanced Linguistic Analyzers* - We provide our custom Lucene analyzers for all aspects of our linguistic analysis. In detail we cover the following linguistic filters:
    - *Capitalization*
    - *Decompounding*
    - *Stemming* - it supports both Solr-provided stemmers and ALiSe's custom stemmer, based on our lingustic resources.
    - *Keyword Protection* - protect any tokens from stemming, handle abbreviations as a single token, etc.
    - *Character Normalization*
    - *NGram support*
    - *Spell checking*
    - *Synonyms*
    - *Taxonomical "Synonyms"*
    - *Word Delimiter / Punctuation*
- *Advanced Linguistic Resources* - Our multilingual resources (dictionary, synonyms, character mappings, endings, etc.) are provided in xml format and are fully customizable. Additionally, the resources themselves can go through customizable analysis during server startup so that they are "lowercased", "stemmed" etc according to the application needs.

5.1. Deployment Options

ALiSe provides two main deployment paths:

- *Single installation* - A pre-built and tuned Solr index configuration is provided. Additionally, a simplified API that can be used to interact with ALiSe is provided.
- *Solr addon* - It can be used upon any existing Solr installations, in the form of a single jar along with the linguistic resources.

5.2. Performance / Scaling

ALiSe is designed to support large volumes of data and handle large amount of queries per second. It is fully distributed. It fully supports SolrCloud technology that provides:

- index distribution across multiple servers
- index replication across multiple servers

It also provides support for legacy Solr scaling through the use of master/server replication.

6. ALiSe Linguistic Resources

ALiSe linguistic text matching is based on language-specific linguistic resources which are editable through the ALiSe Resource Management Tool.

ALiSe Linguistic Resources (per supported language) include:

- **Character Capitalisation Rules**
  It contains the character lower/upper case equivalences for any language.
- **Character Canonization Rules**
  It contains the character canonization equivalences for any language.
- **Functional Words & Phrases**
  It contains the Functional words & Phrases, grouped into compatibility clusters. Each Functional word & Phrase carries also attachment (to Content word) orientation information.
- **Negative Words & Phrases**
  It contains the Negative words & Phrases, grouped into compatibility clusters. Each Negative word & Phrase carries also attachment (to Content word) orientation information.
- **Punctuation**
  It contains the Punctuation patterns, grouped into compatibility clusters. Each Punctuation pattern carries also attachment (to Content word) orientation information.
- **Word Morphological Endings**
  ALiSe uses a proprietary morphological analyser that is based on two levels: the morphological level & the plural level. The word morphological endings of each level are grouped into compatibility clusters.

  For some languages, there is also a “standard” SOLR stemmer which is based on a different set of word morphological endings.

  The ALiSe Morphology module utilizes both options when available.
- **Word & Phrase Synonyms**
  It contains words & phrases grouped into synonym clusters – any morphological variation of a word or phrase is sufficient for the purpose of the Synonym module
- **Taxonomized Terms**
  It contains the taxonomized terms – any morphological variation of a term is

sufficient for the purpose of the taxonomy module – and their relations. Terms are considered "related" based on their distance on the taxonomy.

- **Valid Word Dictionary**

  It contains valid word instances (and their usage counter over big reference data) for the language. It is referenced by the Compounding and Spelling modules

- **Exceptions**

  It contains lists for various types of linguistic resources (e.g. Punctuation, Functional Words & Phrases, etc.) which are exempted from the corresponding SOLR Analyzers. These exemptions can also take the form of a rule (regular expressions).

  For example, the dots inside abbreviations are not regarded as punctuation, certain words take special role in specific contexts (e.g. the word "up" in the phrase "make up" is not regarded as Functional Word), etc.

7. ALiSe Feature Samples

An example of a Query-Result pair-set is provided, which demonstrates several ALiSe features – for simplicity, the example contains textual queries that do not exhibit the range of configurable parameters that ALiSe supports (gaps, strictness, attachment rules, etc.):

Placeholders: the asterisk (*) stands for 0-infinite words; question mark (?) stands for a single word; percent sign (%) stands for 0-infinite characters; and the dot (.) stands for a single character.

Colors: green corresponds to the textual queries, red corresponds to the Boolean queries, blue corresponds to the join queries.

Operator	Intention	Query Expression
	Retrieve all the English Wikipedia article-titles which are	`like (SearchType=FC, Synonyms=ON, TaxoProx=ON):` `anti% animal`
AND	the article has been created later than 2010	
AND		
	the article also contains a sentence-part which is	`like (SearchType=PC):` `anti% resistant bacteria`
OR	contains a whole sentence which is	`like (SearchType=FC+, Synonyms=ON):` `concern medicine may * ? or ?`

The query joins the textual sub-queries on the parent-article-id attribute and returns two results (bold parts refer to content words, underlined parts refer to matched parts):

		Query and Reference Expressions	Comment
Result One	Query	**anti% animal**	
	Result	**Antibiotic** use in **livestock**	*animal* considered as synonym to *livestock*
	… which contains the sentence …		
	Query	**concern medicine may * ?** or **?**	
	Result	Of particular **concern** are **drugs** that **may be passed** in to **milk** or **eggs**	*medicine* considered as synonym to *drug*
Result Two	Query	**anti% animal**	
	Result	**Subtherapeutic antibiotic** use in **swine**	*animal* having a taxo-proximity to *swine*
	… which contains the sentence …		
	Query	**anti% resistant bacteria**	
	Result	**There is concern that use of antibiotics in swine is leading to an increase in resistant bacteria**	

Another example exhibits the compound feature of the textual queries (bold for content words):

	Query and Reference Expressions
Search Type	Full Coverage
Query	**Animal-medicine** for **slaughter farms**
	
Result	**Drugs, administered** to **animals** in **slaughterhouses**