SAP BusinessObjects

Text Data Processing Right-to-Left Language Reference Supplement

- SAP BusinessObjects Data Services 4.1 Support Package 1 (14.1.1.0)

2012-12-20

SAP

Copyright

© 2012 SAP AG. All rights reserved.SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase, Inc. Sybase is an SAP company. Crossgate, m@gic EDDY, B2B 360°, B2B 360° Services are registered trademarks of Crossgate AG in Germany and other countries. Crossgate is an SAP company. All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.These materials are subject to change without notice. These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

2012-12-20

1 Right-to-Left Languages

1.1 Arabic Language Reference

This chapter describes the behavior of the Arabic language module.

1.1.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Arabic texts, including word segmentation, stemming, and tagging.

1.1.1.1 Character Encodings for Arabic

- iso_8859_6
- cp_1256
- utf_8, utf_16, ucs_4

1.1.1.2 Word Segmentation in Arabic

The Arabic segmenter follows all of the general segmentation rules in the white space languages. Specifically, words are broken at whitespace and punctuation. The only exceptions to this are for punctuation embedded inside a word (for example, www.microsoft.command and 1,500), and space embedded in masculine, feminine, or isolated forms of cardinal and ordinal numbers, such as عشرة المائة وأحد.

1.1.1.3 Stemming in Arabic

The major word classes in Arabic stem to the unvocalized stem form, rather than the root form used in traditional Arabic lexicography. This is because too much information is lost when stemming directly to the root. This is shown in the table below. In general, function words stem to themselves. Known proper nouns stem to themselves, and unknown proper nouns are returned unchanged.

Category	Baseform	Examples
Noun	Singular surface form for proper nouns	الجهات -> جهة عمرو -> عمرو
Verb	Verb stem	نأمل -> أمل
Adjective	Masculine singular	الكبيرة -> كبير
Adverb	Adverb stem	هنا -> هنا

1.1.2 Part-of-Speech Tagging in Arabic

The following table shows the Arabic tag set available for defining custom entities. The tag names are accompanied by a brief description and one or more examples. Note that the extraction process recognizes both Arabic and Latin numbers.

Umbrella Tag	Complete Tag	Description	Examples
Abbr		Abbreviation	ان بي سي
Adj	Adj-ConjDet-Fem	Feminine adjective with attached conjunction and determiner	والاقتصادية
	Adj-Det-Fem	Feminine adjective with attached determiner	المتحدة
	Adj-ConjDet-Masc	Masculine adjective with attached conjunction and determiner	والتلفزيوني
	Adj-Det-Masc	Masculine adjective with attached determiner	العقاري
	Adj-Conj-Fem	Feminine adjective with attached conjunction	والارهابية
	Adj-Fem	Feminine adjective	اميركية
	Adj-Conj-Masc	Masculine adjective with attached conjunction	ونهائي
	Adj-Masc	Masculine adjective	احياء
Adv		Adverb	امس
Conj		Detached conjunction	و / أو
Fw		Function word	إن
Interj		Interjection	يا
Nn	Nn-ConjDet-Fem	Feminine noun with attached conjunction and determiner	والدبابات
	Nn-Det-Fem	Feminine noun with attached determiner	الرسالة

Umbrella Tag	Complete Tag	Description	Examples
	Nn-ConjDet-Masc	Masculine noun with attached conjunction and determiner	والمصالح
	Nn-Det-masc	Masculine noun with attached determiner	الأخلاق
	Nn-Conj-Fem	Feminine noun with attached conjunction	فزيارتك
	Nn-Fem	Feminine noun	مخيمات
	Nn-Conj-Masc	Masculine noun with attached conjunction	وتوقيت
	Nn-Masc	Masculine noun	صديقًا
Num	Num	Number	35 الف
	Num-Date	Numeric date	2005/5/23
	Num-Decimal	Decimal number	143.5 100،1
	Num-Time	Numeric time expression	23:00 10:20:11
Part	Part-Fut	Future particle	سوف
	Part-Int	Interrogative particle	هل
	Part-Neg	Negation particle	لما
Prep		Preposition	الى
Pron	Pron-1P	First person pronoun	ونحن
	Pron-2P	Second person pronoun	بك / انت
	Pron-3P-Fem	Third person feminine pronoun	هي / انها
	Pron-3P-Masc	Third person masculine pronoun	هو / فيه
	Pron-3P	Third person pronoun that can be either masculine or feminine	هما / معهما
	Pron-Dem-Fem	Demonstrative feminine pronoun	هذه / وهذي
	Pron-Dem-Masc	Demonstrative masculine pronoun	ذي / هذا

Umbrella Tag	Complete Tag	Description	Examples
	Pron-Rel	Relative pronoun	الذي/ اللتان
Prop	Prop	Proper noun	إنسايت
	Prop-Country	Proper noun that is a country name	اسبانيا / المغرب
	Prop-Email	Email address	info@netscape.com
	Prop-URL	URL address	www.netscape.com
Punct	Punct	Punctuation	؛.،
	Punct-Comma	Comma	، ,
	Punct-Sent	Sentence ending punctuation	.
V	V-Imperativ	Imperative verb	إسع
	V-Imperf	Imperfect verb	تستوجب

The noun and pronoun tags generally do not include number information such as singular, plural or dual. Therefore, the words معلم, معلمين, and معلمان are all tagged as Nn-Masc.

Also, in Arabic, certain prefixes may appear attached to a noun or adjective. The tags for these tokens include information on these prefixes. Therefore, الاتحاد and الاقتصادية will be tagged as Adj-ConjDet-Fem and Nn-Det-Masc, respectively.

Finally, verbs are tagged based on the form used, rather than the actual tense they may represent within the context of the sentence. Therefore, if the verb is formed based on the perfective, it will be tagged as V-Perf, and not as Present Perfect, Simple Past, or Future, for example.

1.1.2.1.1 Unfound Words

Words not found in the tagger dictionary are passed to the Arabic guesser where they are assigned a tag based on a set of rules about Arabic morphology. The following set of tagging rules are part of this module:

- Feminine nominal and adjectival tags are assigned to words ending with a *taa marbuta*. A feminine nominal tag is assigned to words ending with ات.
- A masculine nominal tag is assigned to words ending with ين
- A masculine adjectival tag is assigned to words ending with ي.
- Words that appear to begin with a conjunction, the determiner ال, or both, can also be assigned an adjectival or nominal tag.

- An imperfect verb tag is assigned to words beginning with ت or ي. A perfect verb tag is assigned to words ending with ت (but not ات لمج

Unfound words are also assigned the proper noun tag.

1.1.2.2 Grouping in Arabic

A simple noun phrase in Arabic consists of one of the following:

- Any number of proper nouns optionally followed by any number of definite adjectives **بريطانيا العظمى**
- Any number of definite nouns optionally followed by any number of definite adjectives **القطاع العام**
- Any number of indefinite nouns optionally followed by any number of indefinite adjectives **طالبة جامعية**
- Any number of indefinite nouns followed by any number of definite nouns, and then optionally by any number of definite adjectives **روابط أساتذة الجامعة اللبنانية**

1.1.3 Extraction

This section describes the extraction-specific information for Arabic.

1.1.3.1 Arabic Subtypes

Arabic supports subtypes in the types FACILITY, ORGANIZATION, REGION, URI, COMMON_FACILITY, COMMON_ORGANIZATION, COMMON_PERSON, and COMMON_REGION.

1.1.3.2 Predefined Entity Types

This section describes the predefined entity types supported by the Arabic language module and includes examples of each.

1.1.3.2.1 ADDRESS1

Postal addresses:

38 ش صلاح الدين الاسكندرية

العنوان البريدي ص . ب . 68333 الرياض 11527 المملكة العربية السعودية

1.1.3.2.2 CONTINENT

Any of the continents, for example:

آسيا

أوروبا

افريقيا

Note

America and Australia are extracted as COUNTRY only.

1.1.3.2.3 COUNTRY

The names of countries:

البرازيل

كوريا الجنوبية

اوزبكستان

وطاجيكستان

1.1.3.2.4 CURRENCY

Currency expressions:

2000 درهم

4 دولارات

أربعة دولارات

سبعة وعشرون مارك

1.1.3.2.5 DATE

Dates expressions in the following formats:

1961/3/7

1428/5/12

27 يناير

15 مارس

14 نوفمبر 1995

سبتمبر 2003

1.1.3.2.6 DAY

Names of days of the week:

يوم الجمعة

الأربعاء

يوم الإثنين

1.1.3.2.7 FACILITY

Man-made structures, extracted as one of the following subtypes:

- `AIRPORT`–The names of primarily man-made or man-maintained structures whose primary use is as transportation terminals. For example,
  - ءاسحألا راطم
  - وريديراف راطم
  - ونالي م راطم

- BUILDGROUNDS–The names of architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example,

سفارة المكسيك

المسجد الأقصى

مصرف لبنان

- PATH–The names of primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example,

جسر البحرين

مفرق نابلس

شارع القدس

- PLANT–The names of facilities composed by one or more buildings used for industrial purposes. For example,

مصنع (أيروكونتر)

مفاعل ديمونة

معمل تويوتا

- SUBAREA–The names of portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For example,

ساحة السراي

زنزانة غوانتانامو

1.1.3.2.8 GEO_AREA

Names of geographic regions other than political subdivisions of countries:

- الشرق الأوسط
- الضفة الغربية
- منطقة الخليج

1.1.3.2.9 GEO_FEATURE

Names of places that do not fit into other entity types relating to place (such as Continent, Country, Locality, or Region):

- بحر الشمال
- جزر بحر ايجه
- منطقة جبل بوظهير

1.1.3.2.10 HOLIDAY

Names of holidays:

رأس السنة الهجرية

المولد النبوي

الكريسماس

عيد الفصح

1.1.3.2.11 LOCALITY

The names of cities:

- جوريش
- خربة الرهوة
- طهران
- برلين

1.1.3.2.12 MEASURE

Any measurement, such as weight, volume, or length, in metric units, including standard abbreviations of measurement units:

خمسة أمتار مربعة

ست سنتمترات

14ملم

1.1.3.2.13 MONTH

Names of transliterated Gregorian, Arabic Gregorian, and Hijri months:

إبريل

نيسان

ربيع الثاني

1.1.3.2.14 NOUN_GROUP

Any common noun sequence:

العملية السياسية

المعلومات الوارده

الالعاب الاولمبية

1.1.3.2.15 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The names of commercial organizations, such as major companies or corporations. For example:

إتش إس بي سي

كريدي أغريكول

الراجحي

إسمنت القصيم

شعاع كابيتال

- `MEDIA`–The name of media organizations such as major newspapers or media conglomerates:

وكالةأنباء رويترز

واشنطن بوست

صحيفة الحياة

«الرأي العام»

- `OTHER`--Geopolitical entities that can function as political entities, for example:

الاتحاد الاوروبي

العالم العربي

للمغرب العربي

- `SPORTS`–The name of sports organizations such as major sports franchises or teams:
  - الفيفا
  - نادي الأهلي
  - لباناثينايكوس
  - ريال مدريد

1.1.3.2.16 PEOPLE

Name referring to a group of people based on country, ethnicity, or region:

- آل مكتوم
- بني تميم
- ال نهيان

1.1.3.2.17 PERCENT

Percent expressions:

- %150.5
- 25 بالمائة
- خمس و عشرون بالمائة

1.1.3.2.18 PERSON

Arabic names:

- أم علي
- نواف بن عبد العزيز
- فاروق يوسف
- يحيى أبو زكريا
- طارق محمد الزرقا

Non-Arabic and mixed Arabic/non-Arabic names:

- جورج تينيت
- ورولان عدوان
- لبلير

1.1.3.2.19 PHONE

Telephone numbers:

- +41 22 730 5238
- 360-6130
- 617-782-2669

1.1.3.2.20 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR`--Names of the major political subdivisions of a country, for example:
  - بولاية غوجرات
  - محافظة مأرب
  - وأبين
  - مينيسوتا
- `MINOR`--Names of counties, prefectures, districts, or analogous geographical divisions or governmental units, for example:

تلكلخ

امبابة

الخابورة

1.1.3.2.21 TIME

Expressions of time of day:

- الساعة الحادية عشر
- 11:30

- الخامسة وعشر دقائق

1.1.3.2.22 TIME_PERIOD

Measurements of time, and ranges of time measurements:

- سنتين
- 25 ساعة
- 5 ساعات

1.1.3.2.23 TITLE

Names of titles, positions, or affiliations:

- لعاهل البحرين
- ووكيل وزارة الداخلية اللواء سمو الشيخ
- وزير العدل اللبنانى
- الدكتو

1.1.3.2.24 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`—Email addresses, for example:
  - dot_com@sun.com
- `URL`—Internet addresses, for example:
  - http://www.netscape.com
  - www.netscape.com
  - kcbs.com

1.1.3.3 Common Noun Mentions

Common noun mentions refer to the use of common nouns to refer to entities such as companies, organizations, or cities which would normally also be referred to by proper nouns.

This section describes each common mention entity type supported by the Arabic language module and examples of each.

1.1.3.3.1 COMMON_ADDRESS1

Common names for addresses:

- عنوانه

1.1.3.3.2 COMMON_LOCALITY

Common names for cities:

- العاصمة
- المدن

1.1.3.3.3 COMMON_CONTINENT

Common nouns for the entirety of any continent:

- والقارتين

1.1.3.3.4 COMMON_COUNTRY

Common names for countries and common nouns for political regions that do not fit into more specific common mentions:

- الدولة
- البلدين
- دول
- الأراضي الفلسطينية

1.1.3.3.5 COMMON_FACILITY

Common nouns for man-made structures, extracted as one of the following subtypes:

- `AIRPORTS`—The names of primarily man-made or man-maintained structures whose primary use is as air transportation terminals. For example,
  - مطار
  - المطارات
- `BUILDGROUNDS`—The names of architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example,
  - سفارته
  - المستشفيات
  - المساجد
- `PATH`—The names of primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example,
  - والشوارع
  - مدخل
- `PLANT`—The names of facilities composed by one or more buildings used for industrial purposes. For example,
  - منشأة
  - مصانع
- `SUBAREA`—The names of portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For example,
  - القاعة
  - الغرفة

1.1.3.3.6 COMMON_GEO_AREA

Names of geographic regions other than political subdivisions of countries:

- الشرق الأوسط
- الضفة الغربية
- منطقة الخليج

1.1.3.3.7 COMMON_GEO_FEATURE

Common names for places:

- غابات
- مناطقنا
- المناطق

1.1.3.3.8 COMMON_ORGANIZATION

Common nouns for organizations, extracted as one of the following subtypes:

- `COMMERCIAL`—Common nouns for companies:
  - الشركات
  - الصحيفة
- `OTHER`—Common nouns for organizations that do not fit into a more specific subtype
  - الجمعية
  - المنظمة
  - الغرب

1.1.3.3.9 COMMON_PERSON

Common nouns for persons, extracted as one of the following subtypes:

- `GROUP`—Common nouns for groups of persons
  - المؤلفين
  - اخوان
  - الحشد
- `INDIVIDUAL`—Common nouns for individual persons
  - وعجوز
  - المواطن
  - أقضية

1.1.3.3.10 COMMON_REGION

Common nouns for different regions, extracted as one of the following subtypes:

- `MAJOR`—Common names for states:
  - ولايات
  - المقاطعات
- `MINOR`—Common nouns for the entirety of district areas:
  - فبالقضائين

1.2 Arabic Government Extraction Pack

In addition to extracting all of the entity types found by the standard Arabic module, the Government Extraction Pack has the following behaviors:

- All common mention entity types are extracted by default.
- In addition to the standard entity types, it includes the specialized entity type VEHICLE.
- In addition to the standard common mention types, it includes these specialized common mention entity types: COMMON_VEHICLE, and COMMON_WEAPON.

1.2.1.1.1 VEHICLE

Methods of transportation, extracted as one of the following sub-types:

- `AIR`—Air vehicles, such as airplanes and helicopters. For example,

  وطائرة سنتري

  بوينج 707-320

  التومكات أف-14

  أف 35

- `LAND`—Land vehicles, including the color, year, model and make of the vehicle. For example,

  MERCEDES BENZ

  مرسيدس اس 600

  همر إتش تو

- `WATER`—Water vehicles. For example,

  U.S.S Truxtun

  كول

- `LICENSE`—Alphanumeric sequences that conform to the US and Canadian license plate formats, when preceded by a state abbreviation:

  لوحة رقم 105975

  اللوحه : أ م ق 586

  رقمها /000354 /

- `VIN`—Vehicle Identification Numbers (VIN) in the following format, which always includes 17 characters:

  شاسيه 1248751

  VIN: 2MEFM74W5WX675387

  ورقم الموتور 6510142

  Each position in the VIN has a particular meaning, designating among other things county code, manufacturer code, equipment code, serial number, etc.

1.2.1.1.2 COMMON_VEHICLE

Methods of transportation, extracted as one of the following sub-types:

- AIR—Air vehicles, such as airplanes and helicopters. For example,

طائرات

المروحيات

- LAND—Land vehicles, including the color, year, model and make of the vehicle. For example,

قطار

آلية

وسيارة

- WATER—Water vehicles. For example,

زوارق

سفينتين

المدمرة

1.2.1.1.3 COMMON_WEAPON

Weapons, extracted as one of the following sub-types:

- CHEMICAL—Names of chemical substances that have been officially identified as used to harm humans, plants (crops), and animals, or as potential chemical threats. This also extracts entities that describe the means of the dispersal of any of these weapons.

قنابل الغاز

الاسلحة الكيماوية

- BLUNT— Names of weapons designed or used as bludgeoning instruments.

بالحجارة

هراوة

- SHARP— Names of weapons designed or used to cut, slash, jab, and hack

وسكينا

- EXPLODING—Names of substances that cause damage by exploding.

القنابل

عبوة

- NUCLEAR—Names of weapons that have been officially identified as used to harm humans, plants (crops), and animals through the dispersal of radiological or nuclear energies, or have been identified as potential nuclear threats.

أسلحة نووية

- PROJECTILE—Names of weapons that are designed or used to be projected at great speed for the purpose of causing damage.

والقذائف

صاروخان

برصاص

- SHOOTING—Names of weapons that are designed or used to send projectile objects at great speed for the purpose of causing damage.

بالدبابات

المدفعية

برشاشه

- OTHER—Names of weapons that do not fall into the previous categories

بنيران

ذخيرتهم

سلاح

1.3 Farsi (Persian) Language Reference

This chapter describes the behavior of the Farsi language module.

1.3.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Farsi texts, including word segmentation and stemming.

1.3.1.1 Character Encodings for Farsi

- cp_1256
- utf_8, utf_16, ucs_4

1.3.1.2 Word Segmentation in Farsi

The Farsi segmenter follows all of the general segmentation rules in the white space languages. Certain control characters are also used to mark word boundaries in addition to whitespace and punctuation. For instance, the "zero-width-nonjoiner," which indicates a final form character, is used to delimit word boundaries even when no whitespace is present between the two syntactic units.

Language-specific features are described here.

1.3.1.2.1 Whitespace

Whitespace is not used consistently in Farsi text following a word or punctuation mark. The Farsi segmenter breaks text before and after each punctuation mark, as shown below:

Text	Segmented
فروختند،آنها	فروختند
	,
	آنها

1.3.1.2.2 Final Forms

In Farsi, the final form of letters is used to indicate the end of the word. Final form character markers, such as the 'zero-width joiner' in Unicode and the 'no width non-break' in Windows, are treated like whitespace and are used to segment words. For example:

Text	Segmented
برگ‌درخت‌ها	برگ
	درخت
	ها

However, when a word ends in one of the characters without a final form (such as و ر ز ژ د ذ ا), and there is no whitespace separating it from the following word, then the segmenter does not recognize and segment the word. For instance، افکارمردم and اخباربرگزیده.

1.3.1.2.3 ASCII and Arabic Characters

Some Farsi texts may contain punctuation, numbers and words in ASCII characters, and sometimes Arabic Unicode is used instead of Farsi. The Farsi segmenter is able to recognize and segment ASCII and Arabic Unicode characters.

1.3.1.2.4 Multiword Units

The Farsi segmenter treats multiword units as a single unit.For instance، آسمان‌خراش،گفت و گو. Multiword units may be written with or without whitespace, and the segmenter handles both formats, as shown below:

Text	Segmented
زد و خورد	زد و خورد
زدوخورد	زدوخورد

Hyphens are not included in Farsi multiword units or abbreviation analysis.

1.3.1.2.5 Punctuation

Apart from sentential punctuation (such as period, exclamation point, and question mark), all other punctuation is split from the preceding and/or following word. This is because whitespace between syntactic units or words is optional in Farsi, and is sometimes dropped from text. Also, words like "Yahoo!" do not exist, therefore the exclamation is not kept with the syntactic units.

1.3.1.2.6 Parentheses

Parentheses are split apart regardless of the syntactic unit that appears within the parentheses. There are no such cases in Farsi similar to "fox(es)" or "(re)analyze" in English.

1.3.1.3 Stemming in Farsi

The Farsi stemmer follows the general stemming rules, as described in the Language Reference Guide. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed. Word categories that do not inflect are stemmed to themselves.

Category	Baseform	Text	Stem
Noun	Singular	مسافرین	مسافر
Irregular Noun	Singular	قوانین	قانون
Verb	Infinitive	می‌خندیدند	خندیدن

Adjective	Singular, base form	کوچکترین	کوچک
Adverb	Singular, base form	بیشتر	بیشتر

In Farsi, some inflectional affixes are always attached to the stem, while others are always separated by a final form character or whitespace. Still other affixes may occur either way. The Farsi stemmer removes the affixes uniformly for these variants, as shown below:

Text	Stem
فلسطینیها	فلسطینی
فلسطینی‌ها	فلسطینی
فلسطینی ها	فلسطینی
میروند	رفتن
می روند	رفتن

Cliticized elements and others not belonging to the stem are removed by the stemmer, including clitic pronouns, copulas, and prepositions:

Text	Stem
کتابهایمان	کتاب
وکیلست	وکیل
بشیوه	شیوه
درمنزل	منزل
ازما	ما

The stemmer also removes affixes from deverbal nouns and adjectives and stems these forms to the verb's infinitive form:

Text	Stem
بینندگان	دیدن
بریده	بریدن

1.3.1.3.1 Caveats

In Farsi, some clitics and prefixed prepositions are optionally attached to the stem. The output of the stemmer differs depending on whether these elements are attached or not. For example, if the input is two syntactic units—a preposition followed by a noun, the stemmer outputs two stems. If the preposition is attached to the noun, then the stemmer removes it and returns the noun stem.

Text	Stem
به دفتر	به دفتر
بدفتر	دفتر

Complex verbal constructions are stemmed to their component parts rather than being stemmed as one verbal construction. For instance, both the past tense کار کردند and present tense forms کار می‌کنند of the complex verb کار کردن 'work' are stemmed to the noun کار 'work' followed by the verb کردن 'do'.

If your document contains diacritics or stylistic markers such as short vowels, "tashdid," "sokun," or "keshideh" characters, the syntactic units may not be recognized by the stemmer module.

Note that there is no derivational component in Farsi (Persian) but there is limited derivational stemming in the regular stemmer module for deverbal nouns and adjectives (nouns and adjectives formed from verbs).

1.3.1.3.2 Abbreviations

Abbreviations that end in a period and do not have an intervening space are recognized and tagged as "Abbreviation" by the word-breaking process. For example:

م.

هـق.

Abbreviations with a space or without a period treated in Stemmer. For example:

هـ ق

ق.م

Acronyms are also treated in Stemmer. For example: بی بی سی (BBC)

1.3.2 Part-of-Speech Tagging in Farsi

The following table shows the Farsi tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples.

Umbrella Tag	Complete Tag	Description	Examples
Abbr	Abbr	Abbreviation	هق.
Acc	Acc	Accusative marking on direct objects	را
	Adj-Comp	Comparative singular adjective	بزرگتر
	Adj-NPB	Singular adjective, marks NP boundary	زیبا
	Adj-NoNPB	Singular adjective, not an NP boundary	زیبای
	Adj-Pl	Plural adjective	بزرگان
	Adj-Pl-Comp	Comparative plural adjective	بزرگتران
	Adj-Pl-NPB	Plural adjective, marks NP boundary	زیباها
	Adj-Pl-NoNPB	Plural adjective, not an NP boundary	زیباهای
	Adj-Pl-Sup	Superlative plural adjective	کوچکترینان
	Adj-Sup	Superlative singular adjective	بزرگترین
Adv	Adv	Adverb	همیشه
Conj	Conj	Conjunctions	تا آنجائیکه
Coord	Coord	Coordination	و
Det	Det	Determiner	این
Nn	Nn	Singular noun	کتاب
	Nn-Day	Day noun	دوشنبه

Umbrella Tag	Complete Tag	Description	Examples
	Nn-Mon	Month noun	فروردین
	Nn-NPB	Noun, marks NP boundary	کتابی
	Nn-NonNPB	Noun, not an NP boundary	دانشجوی
	Nn-Pl	Plural noun	کارخانجات
	Nn-Pl-NPB	Plural noun, marks NP boundary	کتاب‌ها
	Nn-Pl-NoNPB	Plural noun, not an NP boundary	کتاب‌های
Num	Num	Number	چهار
	Num-Ord	Ordinal numbers	چهارمین
Phrs	Phrs	Phrase	به به
Prep	Prep	Prepositions	در
Pron	Pron	Pronouns	ما
Prop	Prop	Proper nouns	داریوش
Punct	Punct	Punctuation	،
	Punct-Sent	Sentence boundary punctuation	؟
Qua	Qua	Quantifiers	همه
Que	Que	Question word or wh-phrase	کی
Rel	Rel	Relativizer	که
Title	Title	Titles	آقا
V	V-Fut	Future verb	خواهم رفت
	V-Impv	Imperative verb	برویم
	V-Inf	Infinitive verb	رفتن

Umbrella Tag	Complete Tag	Description	Examples
	V-Part	Past participle or adjectival verb	رفته
	V-Past	Past verb	رفته بودیم
	V-PrPast	Present participle verb	خندان
	V-Pres	Present verb	می‌رویم

1.3.2.1.1 Boundary Markers

The NPB and NoNPB tags signify NounPhraseBoundary and NoNounPhraseBoundary, respectively. These tags are used to indicate whether a boundary defining affix is present on the noun or adjective. Since phrasal boundaries are highly ambiguous in Farsi, this information may be useful for parsing applications that need to distinguish noun phrases and preposition phrases.

Hence, nouns and adjectives ending in the pronoun affix (for example, همسایگانشان 'their neighbors'), the copula verb (مسلمانند 'they are muslims'), the indefinite marker (کتابهایی '(some) books') or the object marker (رئیس‌جمهور را 'the President'), as well as those ending in a vowel (زیبا 'beautiful'), would be tagged with the NPB.

On the other hand, nouns and adjectives ending in an ezafe affix (for example, دانشجوی 'the university student (of)') would be marked with the NoNPB tag since the *ezafe* indicates that the noun should be linked to the following element in the noun phrase.

1.3.2.1.2 Miscellaneous Tags

Nouns that are formed from a verbal root such as شنونده (listener) are tagged as Nouns in the Farsi tagger. Adjectival or participial forms such as شنیده (heard) are tagged as V-Part as shown in the tagset above.

Organization names such as اوپک (OPEC) are tagged as Nouns. Similarly, any foreign word written in ASCII is marked with the Nn tag. However, email and Internet addresses are tagged as Prop (proper names).

Cardinal and ordinal numbers are distinguished in the Farsi tagger since they differ in their syntactic behavior and word order relative to the rest of the noun phrase.

1.3.2.1.3 Unfound Words

The Farsi tagger contains a guesser that determines the tags of the unknown words based on the morphology of the words.

1.3.2.1.4 Grouping in Farsi

Grouping is the recognition of word groups, also called phrases. A phrase is a series of words which form a syntactic unit within a sentence. In Farsi, prepositions are not included in a grouping. The Farsi grouper determines the basic noun phrase elements and groups them together. The simple noun phrase consists of a noun with adjectives, as illustrated by the following:

موشک‌های دوربرد

موسسات خیریه مرتبط

تلاش بین‌المللی

A more complex noun phrase can include several simple noun phrases, consisting of nouns, adjectives, proper names and pronouns. The system makes extensive use of the boundary tags NPB and NoNPB (More Information:

Part-of-Speech Tagging in Farsi [page 22]) in determining the Farsi noun phrases, as illustrated by the following:

وزرای کشور

انسداد اموال القاعده

تلاش‌های گذشته دولت

سازمان‌های پیکارجوی اسلامگرای پاکستان

Note that there are no prepositions or adverbs included in Farsi noun phrase groupings. Determiners, quantifiers, numbers and classifiers are also not grouped as part of the noun phrase.

1.3.3 Extraction

This section describes the extraction-specific information for Farsi.

1.3.3.1 Farsi Subtypes

Farsi supports subtypes in the type ORGANIZATION.

1.3.3.2 Predefined Entity Types

This section describes the predefined entity types supported by the Farsi language module and examples of each: ADDRESS1, COUNTRY, CURRENCY, DATE, DAY, GEO_FEATURE, HOLIDAY, LOCALITY, MEASURE, MISC_NUMERIC, MONTH, NOUN_GROUP, ORGANIZATION, PERCENT, PERSON, PHONE, REGION, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

1.3.3.2.1 ADDRESS1

Street number and street are required parts of the address, city, state, and zip code are optional.

میμان شریعتی

۸۹۹ کوچه‌سوم

1.3.3.2.2 COUNTRY

Names of countries:

افغانستان

آمر یکا

کره‌شمالی

1.3.3.2.3 CURRENCY

The names and quantities of currency expressions:

۴۵۰ دلار

ششصد μ دلار

۸۷ یورو

1.3.3.2.4 DATE

Date expressions:

۸/۸/۱۹۷۵

۳ اردیبهشت ۱۳۷۴

۵ فوریه

1.3.3.2.5 DAY

Days of the week:

جمعه

روز پنجشنبه

1.3.3.2.6 GEO_FEATURE

Any place names other than LOCALITY, REGION@MAJOR or COUNTRY . Includes continents as well as geographic regions and historical sites:

پرسپولیس

آمریکای‌جنوبی

1.3.3.2.7 HOLIDAY

Holidays and special days:

چهارشنبه‌سوری

کریسمس

1.3.3.2.8 LOCALITY

The names of cities:

تهران

واشنگتن

1.3.3.2.9 MEASURE

Measurement, such as weight, volume, or length:

۷۴ متر

هیجده‌هزار گالن

1.3.3.2.10 MISC_NUMERIC

A number together with a person classifier:

۳ نفر

1.3.3.2.11 MONTH

Months of the year:

اردیبهشت

ماه‌فوریه

اسفند ماه

1.3.3.2.12 NOUN_GROUP

One or more nouns with modifying adjectives:

تشدیدخشونت

اعتراض‌دانشجویان

1.3.3.2.13 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- COMMERCIAL– The name of commercial organizations, such as major companies or corporations. For example:

شرکت‌ایران‌خودرو

بانک‌کشاورزی‌ایران

- OTHER–Any organization that does not fit into a more specific subtype.

جبهه مشارکت ایران

حزب‌توده‌ایران

اوپک

1.3.3.2.14 PERCENT

A percentage:

۲۲درصد

1.3.3.2.15 PERSON

An individual specified by name or with a title followed by the name:

داریوش

جورج بوش

آیت‌الله خامنه ای

خانم‌هدایت

کمال‌خارزی

1.3.3.2.16 PHONE

Phone number expressions only when they follow an identifier word such as tel, phone and fax; the identifier is not included in the group:

۸۷۹۷۵۴۵ قغم

۸۸۸۸۸۸ قغر

1.3.3.2.17 REGION

- MAJOR--Names of U.S. States and Iranian provinces, for example:

خراسان

اس a ان خراسان

کالیفرنیا

1.3.3.2.18 TIME

Time expressions:

۱۳ ، ۴۷

۱،۳۰

هفت عصر

1.3.3.2.19 TIME_PERIOD

Measurements of time:

۲۰ دقیقه

دو ساعت

1.3.3.2.20 TITLE

An individual specified by title, or position:

وزیرامورخارجه

معاون بانک ملی ایران

1.3.3.2.21 URI

An address on the internet in ASCII:

- IP – IP addresses, for example:
  - 147.132.42.18
- URL–Internet addresses, for example:
  - http://www.hamshahri.org
  - www.aku.ac.ir

1.3.3.2.22 YEAR

Year and decades:

سال ۸۱

1.4 Hebrew Language Reference

This chapter describes the behavior of the Hebrew language module.

1.4.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Hebrew texts, including word segmentation, stemming, and tagging.

1.4.1.1 Character Encodings for Hebrew

- utf_8, utf_16, ucs_4

1.4.1.2 Word Segmentation in Hebrew

The Hebrew segmenter follows the general segmentation rules for whitespace languages. The Hebrew segmenter has the following language-specific behavior with respect to quotation marks and apostrophes.

A quotation mark " is used as an abbreviation marker in Hebrew, and when found in penultimate position, the " is not separated from the word as a quotation mark. Rather, it is treated as part of the word.

Text	Segmented
ה"ארב	ה"ארב
" לארש'"	"
	לארש'
	"

The apostrophe ' is sometimes used to modify the sound of the characters gimel, tsadi and zayin. This is used in borrowed words where the foreign sound does not match the pronounciation of the Hebrew characters. For example, when the apostrophe occurs to the left of gimel, its sound is changed from the g in good to the j sound in George. In these cases, the apostrophe is segmented together with its word, as shown below.

Text	Segmented
גרו'ג'	גרו'ג'
' לארש '	'
	לארש'
	'

1.4.1.3 Stemming in Hebrew

The Hebrew stemmer follows the general stemming rules. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below.

Category	Baseform	Examples
Noun	singular	השגפ -> ושגפת חול -> תוחול רש -> מרש
Verb	past, 3rd person, masculine, singular form	סנכ -> סנכנ רפס -> רפס
Adjective	masculine, singular form	בוט -> הבוט ונטק -> הנטק רעצ -> תורעצ
Adverb	surface form	לב' -> לב' הפ -> הפ

Clitics are removed from their head words, as shown here:

Text	Stem
קושב	קוש
מאנותעהו	אנותע'
ונצע	עץ
וכתורונמ	הרונמ

1.4.2 Extraction

Note

Hebrew is a basic-level supported language, which means it supports extraction by using dictionaries or extraction rules only.

www.sdn.sap.com/irj/sdn/howtoguides

THE BEST-RUN BUSINESSES RUN SAP™