Knight-Ridder Information's Value Adding Name Finder: A Variation on the Theme of FASTUS

Arkady Borkovsky[1]
(arkady@dnt.dialog.com)

Knight-Ridder Information, Inc., participated in MUC-6 with VANF (Value Adding Name Finder), the system used by Knight-Ridder Information in production for adding a Company Names descriptor field to online newspaper and newswire databases. Knight-Ridder Information participated in the NE task only. The system used for MUC-6 is exactly the same as is used in production. The only difference is the input and output formats. VANF used a cascaded non-deterministic state machines approach and is based on FASTUS.

1. Background

Late 1992, we realized that a company like Knight-Ridder Information (Dialog Information Services, at that time) had to get into the business of Value Added Reselling of information. An obvious question was: what kind of value can be added to the information? The obvious answer was that the added value should have the same nature as the original: it should be information. The plan was to introduce NLP into Dialog's technology to add informational value to raw documents. The industrial environment required that anything we work on has a finite development time and is usable in production after the development is over. This dictated step-wise development and required starting with solvable problems. The obvious choice was *smart tokenization*, which included *named entity detection.*

Originally, we tried to buy the technology from outside, so that a full-fledged partnership could be developed based on the initial project. Two first attempts ended in nothing (although they were quite useful experiences for Dialog, and probably for the potential partners as well). The third one, with SRI, led to development of the current system, VANF (Value Adding Name Finder), which is routinely used in production at Knight-Ridder Information. VANF is an independent re-implementation of FASTUS [1]. At first, we wanted to use FASTUS as it was. However, the production requirements prohibited using a Lisp environment, and we decided to port it (to C++). A direct port was not feasible; at the same time, SRI decided that a declarative, grammar-like description of FSM was more intuitive and easier to work with than the graphical tools they were using at the time. So a grammar definition language was defined, we spent some time discussing and refining it, and I implemented an engine for Dialog's version of the language. Under the contract we had, SRI delivered a transcription of FASTUS rules in the new (declarative) language, although they did not have their own interpreter (nor even a complete language definition) at that time. Somehow, we made this transcription compile and work in Dialog's system, and used it as the core of VANF. Currently, about 30% of VANF rules are derived from FASTUS. Still, VANF should be considered a variation on the theme of FASTUS.

Two thousand documents from different newspapers were processed by human content specialists and the names to be extracted as companies were tagged. The definition of a "company" was "user oriented" (e.g. football teams were to be considered companies, while publications, trade unions and government organizations were not). One thousand of the documents were then used as a training set, and the other 1000 as a blind test set.

[1] Currently with Verity, Inc. reachable as arkady@verity.com

The overall effort on VANF was about 8 months of one developer's time (developing the engine, the lexicons, and the rules), 6 person-months of the content specialists' effort, and about 6 person-months of project management effort.

2. System Description

2.1. Software architecture

VANF consists of

- the Facts Extraction Engine (rules parser and cascaded non-deterministic state machines interpreter)
- the rules set
- evidence combiner
- basic document scanner and lexicon interface
- system interface

The **Facts Extraction Engine** uses cascaded non-deterministic state machines (cascaded NDFSM) to parse the text and to look for patterns.

A grammar defines a set of NDFSM and the sequence of their application.

The terminal symbols in the grammar are literal words and combinations of syntactic and dictionary flags.

Given a NDFSM $\mathcal{M}$ and a sequence of symbols $\mathcal{S}$,

- all the paths are followed;
- the longest matching sequence $\mathcal{S}[1..i]$ is considered the result of the application;
- the actions corresponding to all the longest paths are executed;
- a single output symbol (i.e. a *head word with a bunch of attributes*) is sent to the next stage.

Then $\mathcal{M}$ is applied to the next segment of $\mathcal{S}[i+1..]$.

The output of $\mathcal{M}$ is used as the input for another NDFSM. Each stage is responsible for specific kind of processing (see below).

This approach allows us to restrict non-determinism and prevent combinatorial explosion. The expressive power of our formalism is the same as that of the attribute grammars; at the same time the efficiency is close to that of finite state machines.

The approach is very similar to that used in FASTUS[1].

Evidence combiner and knowledge base. We are not interested in the NDFSM output per se. The pattern matching results are collected from the *side-effect of the actions attached to the rules.*

Most important are the actions which assert facts extracted from the text (in the VANF case, the facts are in the form "*NNN is a name of an entity of type X*"). These assertions are not treated as the ultimate truth, but all the evidence is collected and combined by an evidence combiner.

This includes

- matching the variants of names
- resolving conflicts between contradicting evidence

- checking the names against the known names database

The **document scanner** parses the input documents in their native format (fields, SGML tags, etc.), breaks a document into paragraphs, and does the lexicon look-up. It also takes care of entities like simple dates, phone numbers, and other regular quasi-lexical units. This part of the scanner is based on regular expressions (`lex`).

2.2. Text processing flow

1) The basic tokenizer (written in lex) handles lexicon lookup, capitalization, simple money, number, phone-number, and date expression processing, simple unknown names processing; multi-word lexicon entries are handled here as well.
2) The preprocessor (1st stage of NDFSM) looks for sentence boundaries, more complicated number and date expressions, obvious names type determination and other names bracketing.
3) The chunker (2nd stage of NDFSM) brackets noun phrases and verb groups, replacing them with their head words.
4) The pattern matchers; several stages of NDFSM detect certain interesting types of noun groups using patterns("*ABC's president X*", and look for other patterns relevant to the VANF task; for each interesting pattern an attached function is executed to pass on a candidate name to the evidence combiner.
5) The evidence combiner (C++) merges similar names and assigns types to the names whose type could not be derived from the form or from the context; database lookup happens here as well.

Thus, the name type definition is performed based on several kinds of evidence:

- name's form (during step 2)
- name's context (during step 4)
- similarity with already defined name (step 5)
- database lookup (step 5)

3. The History and the Future

3.1. "Where the bulk of effort was spent"

Knight-Ridder Information, Inc. did only the NE part.

The original VANF development tool:

4 months - the engine
4 months - the rules and improving the lexicon

Then, for the MUC-6 evaluation:

1 month - preparation for MUC; this included improving the rules based on the training corpus and making the system interface component to produce output, which made the MUC scorer happy.

3.2. Technical characteristics

Processing rate: 5 Mb / hour
Reloading the lexicon: 20 sec
Reloading the grammar: 20 sec

3.3. Training and rules modification

All rules development was done by hand. The engine provides a lot of tracing information and allows easy definition of which rule(s) contributed to a specific decision. This should help to implement automatic training (see below) in the future. At present, this allows us to modify rules. Also, the reason we could get anywhere at all was the fast turn-around cycle: modifying the grammar, and rerunning a document takes about 1 minute.

3.4. Experience gained from VANF and the future

Our experience with VANF has proved that a core cascaded NDFSM approach is suitable for many intelligent text processing tasks. As it was pointed out in [1], an efficient implementation and easy-to-use tools allowed us to tune a pretty primitive technology to produce quite useful results. In our case, such efficiency made it possible to build a rule set consisting of many quite specific rules, such that although each one has a limited application, together, they cover a large area. In the future, such rule sets should be constructed using automated tools. Two future tasks will be concentrated on:

- training / rules building tools; the author plans to develop a learn-by-example system, conceptually similar the Autoslog [2] and the like;
- non-boolean evidence combining.

4. Appendix 1: Lexicon and Grammar Samples

The lexicon contains word forms with flags; the flags encode POS and semantic information. There is no formal differentiation between primary (N, V) and secondary (Sing, Trans) morphological features, nor between grammatical and semantic flags. The flags corresponding to different meanings of a word are mixed in the same entry. (Therefore, in the grammar, N[Country] or Country[N] are equivalent notations. One can also write N[V] meaning "any word which is both a verb and a noun".)

4.1. A sample of the lexicon

```
aging  N Sing V-Ing Trans
agitate  V Trans
agitated  V-Ed-En Trans
agitates  V-S Trans
Argentine  Adj Noun-Like
Argentine  N Sing Adjnoun
Argentinean  N Sing Adjnoun Nationality
Argentineans  N-S Nationality
Arizona  N Sing Country State
'so called'  Adj
'so far'  Adv
'so that'  Subconj
'soft drink'  N Sing
'soft drinks'  N-S
'software house'  N Sing
```

Lexical information can be also stored in the grammar description in the form of word lists.

```
$relative$ -->
        'mother'  'father'
        'parents'  'grand-mother'  'grand-father'  'grand-parents'
        'son'      'sons'      'daughter'  'daughters'   'child'
        'children' 'grand-daughter' 'grand-son'
        'wife'     'ex-wife'   'husband'    'ex-husband'  'spouse'
        'sister'   'sisters'   'brother'    'brothers'     'sibling'   'siblings'
        ;;
```

4.2. A sample of the grammar

For example, the "syntax stage" of the VANF NDFSM contains rules :

```
VG ==> .... | Be-VG | ....
Be-VG --> Aux-Be-VG  Adv*
        ; Be = T
        ;;
        % am not, is not, etc. as main verb.
        % ain't, isn't, etc. as main verb.
        % will often be, can't have been, has been, had not been
Aux-Be-VG --> {         Ax-Be
                |       Aux-Modal Adv* 'be'
                |       Aux-Have 'been'
                }
                ;;
        % "is [broken]' 'isn't [really broken]', 'will often be [broken]',
        % 'can't have been [broken]', 'has been [broken]', 'had not been [broken]."
Aux-Be --> { Bel | Be-Not | Aux-Modal 'be' | Aux-Have 'been' } ('being') ;;
Bel --> { 'am' | 'is' << Sing = T >>| 'are' | ''re' | 'was' << Sing = T >>| 'were'
        } ;;
Be-Not --> { Bel 'not' | 'ain't' | 'aren't' | 'isn't' << Sing = T >> | 'wasn't' <<
        Sing = T >> | 'weren't' }
        ; neg = T
        ;;
```

The following rule belongs to a sub-NDFSM which consumes the output of the "syntax phase" and detects person names in contexts like "*Neither of Makoto Suzuki's parents*" or "*Mary was Joe's wife*":

```
Name-In-Context --> (Possible-Person-Name (',' | Be)) { Possible-Person-Name | NG
        } ''s' NG[-Indef,$relative$]  ;;
```

5. Appendix. 2. The Walk-Through Example

5.1. Errors list.

1) Punctuation -
2) "*Interpublic Group's McCann-Erickson*" and "*WPP Group's J. Walter Thompson*" were treated as single entities; this is strange, because we have rules to split them, and the evidence combiner must have picked the second parts of these names, because they occur elsewhere. A possible explanation is that Dialog's original requirements explicitly advised against splitting the names, which might make sense in the first of the two examples.
3) "*60 pounds*" taken for a monetary expression.
4) All occurrences of "*Coke*" were ignored.
5) "*New York Times*" ignored - according to Dialog's spec.
6) The Title field was not processed at all - for no good reason.
7) "Other ad agencies, such as <ENAMEX TYPE="PERSON">Fallon McElligott</ENAMEX>," - a bug

5.2. The output

```
<DOCID> wsj94_026.0231 </DOCID>
<DOCNO> 940224-0133. </DOCNO>
<HL>    Marketing & Media -- Advertising:
@  John Dooner Will Succeed James
@  At Helm of McCann-Erickson
@  ----
@  By Kevin Goldman </HL>
<DD> <TIMEX TYPE='DATE'>02/24/94</TIMEX> </DD>
<SO> WALL STREET JOURNAL (J), PAGE B8 </SO>
<CO>    IPG K </CO>
<IN> ADVERTISING (ADV), ALL ENTERTAINMENT & LEISURE (ENT),
   FOOD PRODUCTS (FOD), FOOD PRODUCERS, EXCLUDING FISHING (OFP),
   RECREATIONAL PRODUCTS & SERVICES (REC), TOYS (TMF) </IN>
<TXT>
<p>
   One of the many differences between <ENAMEX TYPE='PERSON'>Robert L.
James</ENAMEX>, chairman and
chief executive officer of <ENAMEX TYPE='ORGANIZATION'>McCann-Erickson</ENAMEX>, and
<ENAMEX TYPE='PERSON'>John J. Dooner Jr</ENAMEX>.,
the agency's president and chief operating officer, is quite
telling: Mr. <ENAMEX TYPE='PERSON'>James</ENAMEX> enjoys sailboating, while Mr.
<ENAMEX TYPE='PERSON'>Dooner</ENAMEX> owns a
powerboat.
</p>
<p>
   Now, Mr. <ENAMEX TYPE='PERSON'>James</ENAMEX> is preparing to sail into the
sunset, and Mr.
<ENAMEX TYPE='PERSON'>Dooner</ENAMEX> is poised to rev up the engines to guide
<ENAMEX TYPE='ORGANIZATION'>Interpublic Group's
McCann-Erickson</ENAMEX> into the 21st century. Yesterday, <ENAMEX
TYPE='ORGANIZATION'>McCann</ENAMEX> made
official what had been widely anticipated: Mr. <ENAMEX TYPE='PERSON'>James</ENAMEX>,
57 years old,
is stepping down as chief executive officer on <TIMEX TYPE='DATE'>July 1</TIMEX> and
will
retire as chairman at the end of the year. He will be succeeded by
Mr. <ENAMEX TYPE='PERSON'>Dooner</ENAMEX>, 45.
</p>
<p>
   It promises to be a smooth process, which is unusual given the
volatile atmosphere of the advertising business. But Mr. <ENAMEX
TYPE='PERSON'>Dooner</ENAMEX> has
a big challenge that will be his top priority. 'I'm going to focus
on strengthening the creative work,' he says. 'There is room to
grow. We can make further improvements in terms of the perception of
our creative work.'
```

```
</p>
<p>
   Even <ENAMEX TYPE="PERSON">Alan Gottesman</ENAMEX>, an analyst with <ENAMEX
TYPE="ORGANIZATION">PaineWebber</ENAMEX>, who believes
<ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> is filled with "vitality" and is in
"great shape," says that
from a creative standpoint, "You wouldn't pay to see their reel" of
commercials.
</p>
<p>
   While McCann's world-wide billings rose <NUMEX TYPE="PERCENT">12%</NUMEX> to
<NUMEX TYPE="MONEY">$6.4 billion</NUMEX> last
year from <NUMEX TYPE="MONEY">$5.7 billion</NUMEX> in <TIMEX
TYPE="DATE">1992</TIMEX>, the agency still is dogged by the
loss of the key creative assignment for the prestigious Coca-Cola
Classic account. "I would be less than honest to say I'm not
disappointed not to be able to claim creative leadership for Coke,"
Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> says.
</p>
<p>
   <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> still handles promotions and media
buying for Coke. But
the bragging rights to Coke's ubiquitous advertising belongs to
<ENAMEX TYPE="ORGANIZATION">Creative Artists Agency</ENAMEX>, the big <ENAMEX
TYPE="LOCATION">Hollywood</ENAMEX> talent agency. "We are
striving to have a strong renewed creative partnership with
Coca-Cola," Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> says. However, odds of that
happening are
slim since word from Coke headquarters in <ENAMEX TYPE="LOCATION">Atlanta</ENAMEX>
is that <ENAMEX TYPE="ORGANIZATION">CAA</ENAMEX> and
other ad agencies, such as <ENAMEX TYPE="PERSON">Fallon McElligott</ENAMEX>, will
continue to
handle Coke advertising.
</p>
<p>
   Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX>, who recently lost <NUMEX
TYPE="MONEY">60 pounds</NUMEX> over three-and-a-half
months, says now that he has "reinvented" himself, he wants to do
the same for the agency. For Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX>, it means
maintaining his
running and exercise schedule, and for the agency, it means
developing more global campaigns that nonetheless reflect local
cultures. One <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> account, "I Can't Believe
It's Not Butter," a
butter substitute, is in 11 countries, for example.
</p>
<p>
   <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> has initiated a new so-called global
collaborative system,
composed of world-wide account directors paired with creative
partners. In addition, <ENAMEX TYPE="PERSON">Peter Kim</ENAMEX> was hired from
<ENAMEX TYPE="ORGANIZATION">WPP Group's J.
Walter Thompson</ENAMEX> last <TIMEX TYPE="DATE">September</TIMEX> as vice chairman,
chief strategy
officer, world-wide.
</p>
<p>
   Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> doesn't see a creative malaise
permeating the agency.
He points to several campaigns with pride, including the Taster's
Choice commercials that are like a running soap opera. "It's a <NUMEX
TYPE="MONEY">$19
</NUMEX>million campaign with the recognition of a <NUMEX TYPE="MONEY">$200
million</NUMEX> campaign,"
he says of the commercials that feature a couple that must hold a
record for the length of time dating before kissing.
</p>
<p>
   Even so, Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> is on the prowl for more
creative talent and
is interested in acquiring a hot agency. He says he would like to
finalize an acquisition "yesterday. I'm not known for patience."
</p>
<p>
```

```
   Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive
officer of <ENAMEX TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency
with billings of <NUMEX TYPE="MONEY">$400 million</NUMEX>, but nothing has
materialized. "There
is no question," says Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX>, "that we are
looking for quality
acquisitions and <ENAMEX TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX> is a quality
operation. There are
some people and entire agencies that I would love to see be part of
the <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> family." Mr. <ENAMEX
TYPE="PERSON">Dooner</ENAMEX> declines to identify possible
acquisitions.
</p>
<p>
   Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> is just gearing up for the headaches of
running one of
the largest world-wide agencies. (There are no immediate plans to
replace Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> as president; Mr. <ENAMEX
TYPE="PERSON">James</ENAMEX> operated as chairman,
chief executive officer and president for a period of time.) Mr.
<ENAMEX TYPE="PERSON">James</ENAMEX> is filled with thoughts of enjoying his three
hobbies:
sailing, skiing and hunting.
</p>
<p>
   Asked why he would choose to voluntarily exit while he still is
so young, Mr. <ENAMEX TYPE="PERSON">James</ENAMEX> says it is time to be a tad
selfish about how he
spends his days. Mr. <ENAMEX TYPE="PERSON">James</ENAMEX>, who has a reputation as
an
extraordinarily tough taskmaster, says that because he "had a great
time" in advertising," he doesn't want to "talk about the
disappointments." In fact, when he is asked his opinion of the new
batch of Coke ads from <ENAMEX TYPE="ORGANIZATION">CAA</ENAMEX>, Mr. <ENAMEX
TYPE="PERSON">James</ENAMEX> places his hands over his
mouth. He shrugs. He doesn't utter a word. He has, he says, fond
memories of working with Coke executives. "Coke has given us great
highs," says Mr. <ENAMEX TYPE="PERSON">James</ENAMEX>, sitting in his plush office,
filled with
photographs of sailing as well as huge models of, among other
things, a Dutch tugboat.
</p>
<p>
   He says he feels a "great sense of accomplishment." In 36
countries, <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX> is ranked in the top three;
in 75 countries, it is
in the top 10.
</p>
<p>
   Soon, Mr. <ENAMEX TYPE="PERSON">James</ENAMEX> will be able to compete in as many
sailing races
as he chooses. And concentrate on his duties as rear commodore at
the <ENAMEX TYPE="ORGANIZATION">New York Yacht Club</ENAMEX>.
</p>
<p>
   Maybe he'll even leave something from his office for Mr. <ENAMEX
TYPE="PERSON">Dooner</ENAMEX>.
Perhaps a framed page from the <ENAMEX TYPE="ORGANIZATION">New York Times</ENAMEX>,
dated <TIMEX TYPE="DATE">Dec. 8, 1987</TIMEX>,
showing a year-end chart of the stock market crash earlier that
year. Mr. <ENAMEX TYPE="PERSON">James</ENAMEX> says he framed it and kept it by his
desk as a
"personal reminder. It can all be gone like that."
</p>
</TXT>
</DOC>
```

6. References

[1] Douglas E. Appelt, Jerry R. Hobbs , David Israel, Mabry Tyson "FASTUS: A Finite-state Processor for Information Extraction from Real-world Text". In IJCAI 93, 13 Joint Conference of Artificial Intelligence.

[2] Riloff, E. "Automatically Constructing a Dictionary for Information Extraction Tasks" Proceedings of the 11th National Conference on Artificial Intelligence, 811-816.