SALE - Software Architecture for Language Engineering

Language Engineering: „Production of software systems that invorlve processing human language with quantifiable accuracy and predictable development resources“ (Cunningham, 1999)

Die Herstellung von Software-Systemen, die *die Verarbeitung natürlicher Sprache mit quantitativ bestimmbarer/quantifizierbarer Präzision* sowie *voraussagbare Entwicklungsressourcen* beinhalten.

LE ist verwandt zu den Bereichen CL, NLP und AI, hat aber einen von ihnen verschiedenem Focus und setzt anderen Prioritäten. Die wichtigsten Ziele und Motivationen sind:

- Behandlung von Aufgaben praktischen Nutzens in großem Maßstab.
- Messung des Fortschritts in quantitativer Relation zu der Effizienz / Performanz anhand von Beispielen solcher Aufgaben.
- Wachsende Wichtigkeit von Software Engineering im Allgemeinen
- Wiederverwendbarkeit
- Robustheit
- Effizienz
- Produktivität

SALE
Konstruktion einer softwaretechnischen Infrastruktur für die Verarbeitung von Sprache.
Entspricht einem „Werkzeugkasten“ zur Konstruktion von Systemen und Experimenten.
3 Typen von infrastrukturellen Systemen: Frameworks, Architekturen und Entwicklungsumgebungen.

Framework (Gerüst)/ Platform / Component System: Objektorientierte Klassenbibliothek, die für einen bestimmten Bereich konstruiert wurde und für die Lösung von Problemen in diesem Bereich erweitert oder zugeschnitten werden kann.

Alle Software-Systeme haben eine explizite oder implizite Architektur.
Explizit (Reference Architecture / Domain-Specific-Architecture): Evtl. allg. Standards entsprechend, auf mehrere Systeme gerichtet
Implizit (Software architecture fon a family of application systems)

Eine Implementation einer Architektur, die z. B. graphische Werkzeuge für die Erstellung und das Testen von Systemen bereitstellt, ist eine Entwicklungsumgebung.

Stichpunkte zu SALE systems:

1. Strikte Trennung von systemnahen Aufgaben wie Datenspeicherung, Datenvisualisierung, Speicherstelle und Laden von Komponenten und der Ausführung von Prozessen von den Datenstrukturen und Algrorithmen, die tatsächlich die natürliche Sprache verarbeiten.
2. Die Reduzierung von Integrationskosten durch die Bereitstellung von standardisierten Mechanismen für die Sprachdaten-Übertragung zwischen Komponenten und die Verwendung offener Standards wie Java und XMLals zugrundeliegende Plattform.
3. Bereitstellung eines Basis-Sets von evtl. Komponenten zur Sprachverarbeitung, dessen Komponenten ggf. vom Benutzer (z. B. zu vergleichszwecken) ausgetauscht werden können und welches erweitert werden kann.
4. Bereitstellung einer Entwicklungsumgebung oder zumindest eines Sets von Werkzeugen zur Unterstützung des Benutzers bei der Modifikation und Implementation von sprachverarbeitenden Komponenten und Anwendungen.
5. Automatische Messung der Performanz der sprachverarbeitenden Komponenten.

Work on SALE:
LE-Programme bestehen aus Daten und Algorithmen.
Trend in der Software-Entwicklung: Daten und Algorithmen zusammen als Objekte modellieren, weil leichter zu erstellen und zu pflegen
Bei der Verarbeitung natürlicher Sprache ist eine strikte Trennung nicht möglich. Sprachdaten haben eine solche Aussagekraft, dass sie häufig unabhängig von Algorithmen bearbeitet werden.
(???) --> Language Resources (Lexika, Korpora)

Language Resource (LR): Reine Datenressourcen wie Lexika, Korpora, Thesauri, Ontologien. Zu manchen LRs wird Software bereitgestellt, doch da diese nur ein Mittel zum Datenzugriff darstellt, werden auch diese Ressourcen als LRs angesehen.

Processing Resource (PR): Prinzipiell programmatischer, algorithmischer Charakter, wie z. B. bei Lemmatisierern, Generatoren, Übersetzern, Parsern oder Spracherkennern.
Bsp.: POS-Tagger ist am besten charakterisierbar duch die Prozessierung, die er mit dem Text durchführt. PRs enthalten oft LRs, z. B. enthalten Tagger häufig Lexika.
PRs: Algorithmen, die verschiedene Typen von LRs aufeinander abbilden (mappen) und die dazu LRs verwenden. Ein Programm zur MÜ z. B., bildet einen einsprachigen Korpus auf einen mehrsprachig ausgerichteten ab, der Lexika, Grammatiken etc. verwendet.
Die Übernahme der LR/PR-Unterscheidung ist eine Art, der bestehenden Praxis und Terminologie zu entsprechen. Sie widerspricht keineswegs dem Prinzip der Objektorientierung.
Unterscheidung der Begriffe, um die Verarbeitung mit einer SALE zu kategorisieren.